Computationally efficient statistical approaches for spatial modelling and high-dimensional emulation of tsunami models

Xiaoyu Liu

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Statistical Science

University College London

November 28, 2016

I, Xiaoyu Liu, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This thesis addresses some issues in quantifying spatial uncertainties and their propagation through computer models using statistical emulation, motivated by the uncertainty quantification of bathymetry for tsunami modelling. Firstly, we develop a computationally efficient model for spatial data. Gaussian fields (GFs) are frequently used but the associated computational cost can be a bottleneck. The efficient SPDE approach has been proposed by doing the computations using Gaussian Markov random fields (GMRFs) as GFs can be seen as weak solutions to the corresponding stochastic partial differential equations (SPDEs) using piecewise linear finite elements. We introduce a new class of representations of GFs with bivariate splines instead of finite elements. This allows an easier implementation of piecewise polynomial representations of various degrees. It leads to GMRFs that can be inferred efficiently and can be easily extended to non-stationary fields. Secondly, we build statistical emulation for computer models with high-dimensional inputs. In this case, the construction of an emulator can become prohibitively expensive. We propose a joint framework merging emulation with dimension reduction in order to overcome this hurdle. The gradient-based kernel dimension reduction method is chosen for its ability to extract drastically lower dimensions with little loss in information. This generates a low-dimensional process which is emulated with a Gaussian process. The proposed framework is demonstrated to be effective and efficient both theoretically and numerically. Finally, we consider the geostatistical inference of multiple spatial surveys that usually differ in aspects like resolution, accuracy and location. Geoscientific surveys sometimes also present preferential sampling features, which suggest that data locations depend on the values of the

spatial field. We propose a joint hierarchical model based on the SPDE approach. This joint model allows us to account for the respective characteristics in each of the surveys separately and thus makes the inference for the underlying spatial process more accurate.

Acknowledgements

First and foremost, I would like to sincerely thank my primary supervisor Prof. Serge Guillas. He opened the door to this interesting interdisciplinary field and guided me to learn and do research. With his enduring support, I am able to make some contributions, which might be small in others' eyes but mean a lot to me. Throughout my PhD programme, he sought and provided me a lot of opportunities such as building collaboration and attending conferences. I have learned a lot from him which would be helpful in my whole life.

I am also grateful to my second supervisor, Dr. Ioanna Manolopoulou. She gave me a lot of constructive suggestions in the whole process, especially for my PhD upgrade, for which I would like to acknowledge Dr. Giampiero Marra as well. I would also like to thank Prof. Ming-Jun Lai for his help in bivariate spline techniques and the collaboration which led to my first publication during the programme.

I would like to acknowledge the China Scholarship Council and UCL for providing me financial support to take this long journey in London. The same gratitude also goes to the UCL Advances for the support to evaluate the commercial value of my research findings. I am grateful to the staff and colleagues in the Department of Statistical Science for making the department such a supportive, collaborative and energetic environment. I acknowledge that some of the work in this thesis made use of Emerald, a GPU-accelerated High Performance Computer, made available by the Science & Engineering South Consortium operated in partnership with the STFC Rutherford-Appleton Laboratory. I also acknowledge the use of the UCL Legion High Performance Computing Facility (Legion@UCL), and associated support ser-

vices, in the completion of this work.

I want to also thank my future girlfriend who has kept invisible so far to let me focus on the study. Most importantly, I would like to express my endless gratitude and love to my family. They are always my motivation to fight for a better myself!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and motivation

A tsunami is a series of water waves that are triggered by the underwater disturbance associated with primarily earthquakes or landslides. It is one of the most dangerous natural hazards. The devastating tsunami disaster that occurred on the Boxing Day of 2004 at the Island of Sumatra in Indonesia hit the coast with waves up to 30 metres high and killed about 230,000 people. A more recent tsunami event combined with an earthquake attacked Tōhoku in Japan on March 11, 2011 leading to nearly 20,000 deaths. These tragedies have raised people's awareness of the importance of tsunami research and the associated risk assessment.

In general, the whole life-cycle of a tsunami can be described as three stages: generation, propagation and inundation. The principal generation mechanism of a tsunami is the displacement of a substantial volume of water. The displacement could be generated by submarine or coastal earthquakes and landslides, volcanoes, meteorological activities and human-caused explosions. For a comprehensive summary and comparison of the tsunami sources, please refer to Sarri (2015) and the references therein. After being generated, the tsunami waves propagate as long waves. Compared with usual waves, the wavelength of a tsunami is generally larger and the amplitude is smaller. The waves travel at high speed and carry high energy to a long distance. When approaching the coast, the waves are amplified and the speed and wavelength are decreased. The large volume of water flow can inundate

the coastal region causing tremendous damage to infrastructures and human lives.

Tsunamis are among the extremely rare hazards, it is difficult to gather sufficient information from past events to study them. Computer models, or namely simulators, are developed and applied with mathematical and geophysical knowledge to simulate tsunami scenarios. These simulations can be used to produce hazard maps or in early warning systems to help mitigate the hazard. They are also frequently applied in insurance industry to help with the assessment of potential losses using catastrophe models; see Appendix C for more details. There are several operational numerical models available for tsunami simulations, such as MOST, NAMI, ComCot (Goto et al., 1997, Imamura, 1996, Liu et al., 1998, Titov and Gonzalez, 1997). However, the numerical schemes embedded in these codes are essentially out of date. VOLNA (Dutykh et al., 2011) is one of the most recently developed tsunami codes, which takes both advantages of modern numerical techniques and high computing power. In this thesis, the VOLNA code is employed to handle the whole life-cycle of tsunamis. VOLNA has been implemented with high performance computing techniques including GPU and parallel computing on the GPU cluster Emerald by the team led by Prof. Mike Giles and Dr. Istvan Reguly at the University of Oxford. This makes the evaluation much faster by a factor of 3000 to 8000 compared to the serial version on a normal desktop. The VOLNA code has been applied to investigate the tsunami risk over Cascadia region (Sarri, 2015); see Appendix B for some practical considerations about the simulation including bathymetry data, triangulation construction and examples of coastal hazard maps.

Despite the wide use of computer models in complex natural or societal phenomena, they only mimic or approximate the real-world processes. Sometimes, even the most advanced models cannot represent the reality exactly, or some model components cannot be known exactly or are random in nature. Therefore, it is necessary to conduct uncertainty quantification (UQ), which aims to characterise the model behaviour when uncertainties are involved, to evaluate the model performance and produce more reliable predictions and analysis. This is especially crucial for tsunami research due to the fact that there are a lot of unknowns and uncertain-

ties in the complicated processes. For example, González et al. (2009) took into account several uncertain factors to develop a probabilistic tsunami flooding map using the tsunami inundation code MOST integrated with methods of probabilistic seismic hazard assessment. They considered the possible multiple earthquake sources and several causes of uncertainty such as the tidal stage at tsunami arrival, near-field slip distribution and inter-event time.

Tsunami simulators, just like many other complex computer models, are often computationally demanding. It makes the uncertainty quantification expensive or even prohibitive since sufficient number of simulations are not affordable, e.g. using conventional Monte Carlo method. In this case, a computationally efficient statistical surrogate, known as emulator, is often constructed to replace the expensive simulator. It is usually able to make accurate predictions with only a small number of well-designed simulations, and carry out expensive tasks such as uncertainty and sensitivity analysis. It has been successfully applied to tsunami research. For example, Sarri et al. (2012) discussed uncertainties in the sources of landslide generated tsunami such as the position, shape and speed of a landslide. By using the Gaussian process techniques, they built a fast statistical emulator of the landslide generated tsunami computer model for efficient uncertainty quantification and sensitivity analysis. A more recent study by Sraj et al. (2014) investigated the impact of Manning's n friction coefficient, which represents the effect of bottom friction, on the tsunami wave elevations. The polynomial chaos emulator was applied for uncertainty propagation and sensitivity analysis. They further estimated and quantified the uncertainty in the Manning's n coefficients using Bayesian methods with data collected during the 2011 Tōhoku tsunami event.

Most of the current research in the uncertainty quantification of tsunamis are focused on the generation or physical parameters. However, the impact of uncertainties in the bathymetry (a metric of the seafloor elevation) has not been addressed sufficiently in the community. The effect of the seafloor characterisation on tsunami waves has been noticed. Iglesias et al. (2014) investigated the variations in tsunami propagation and hence the impact over the coast because of the presence of a subma-

rine canyon incised in the continental margin. Their simulation results revealed the significant effect of the presence, morphology and orientation of submarine canyons on the arrival times and amplitudes of a tsunami. This has highlighted the need for precise seafloor characterisation as well as proper treatment to the uncertainties for tsunami modelling. Moreover, precise land elevations (topography) are also critical to accurate tsunami inundation calculations. In this thesis, we may refer to both bathymetry and topography when mentioning "bathymetry" where appropriate.

Eakins and Taylor (2010) introduced the general procedure to produce an integrated bathymetric and topographic digital elevation model (DEM) at the National Oceanic and Atmospheric Administration (NOAA). They asserted that building high-resolution, integrated bathymetric and topographic DEMs are essential for tsunami modelling. Elevation data from multiple sources are gathered and converted to a common file format and reference frame first. Then, these raw data are mapped to grids using the Generic Mapping Tools (GMT) (Wessel and Smith, 1998) or some other similar gridding systems. The ideas behind GMT can be found in Smith and Wessel (1990) and Wessel and Bercovici (1998). There are some other bathymetric data products such as the gridded bathymetry data (GBD) by the General Bathymetric Chart of the Oceans (GEBCO) (Hall, 2006). However, most of the current bathymetric data products are generated as grids which are possibly not appropriate for the newly developed tsunami models, e.g. VOLNA, that employ advanced numerical schemes with irregular and unstructured mesh. These products usually do not include uncertainty estimates so that it is not straightforward to assess their impact on the outputs of tsunami models.

The impact of the uncertainties in the DEMs has been investigated on a geophysical flow model of volcano when more than one DEMs are available for the same geographical region. Stefanescu et al. (2012b) illustrated that DEMs of different resolutions and sources could lead to different outputs and hence different flow maps. It was concluded that fine DEM resolution is critical to correctly characterise the granular flows. However, some high-resolution DEMs are created by just decreasing the interval between grid points in the interpolation. Stefanescu et al.

(2012b) showed through several numerical simulations that interpolation might result in a measurable negative effect on the model outputs. Stefanescu et al. (2012a) proposed two methods to quantify the uncertainties in the DEMs through the so-called error map which is the difference between two DEMs of different resolutions for the same area. One method assumes that the errors are spatially uncorrelated while the other adopts an autocorrelation structure. The uncertainties were propagated through the geophysical flow model using emulation with the Bayes linear method (Goldstein and Wooff, 2007). The outputs were appropriately combined to produce probabilistic hazard maps. The results showed that it is critical to consider the spatial autocorrelation structure in order to incorporate the uncertainties in the DEMs properly.

There have been new interests in dealing with complex spatial data and the uncertainties. For example, Sangalli et al. (2013) proposed the spatial spline regression (SSR) model to analyse data distributed over irregular domains. This model is able to easily handle complex boundary conditions, concavities and interior holes. It provides a large advantage over the other classical techniques such as kriging and thin-plate splines when dealing with data scattered over irregularly shaped domains. In another innovative approach, Lindgren et al. (2011) considered the latent Gaussian models where the latent Gaussian fields (GFs) can be explicitly linked to Gaussian Markov random fields (GMRFs) through the associated stochastic partial differential equations (SPDEs). The computation is carried out on the GMRFs instead of GFs which makes the full Bayesian inference much more efficient using the integrated nested Laplace approximations (INLA) by Rue et al. (2009). The SPDE approach can be applied to make predictions of a spatial process at any locations by computing their posterior distributions given the observations. Sangalli et al. (2013) pointed out that their SSR model has strong connections with the work of Lindgren et al. (2011). The statistical nature of the SSR and SPDE approaches as well as other geostatistical models provides us a direct way to handle the uncertainties in the spatial data.

Therefore, motivated by tsunami research, this thesis addresses some issues in

two problems: (1) how to quantify the uncertainties in the bathymetry; (2) how to propagate these uncertainties to tsunami waves. We attack the first problem based on the novel SPDE approach with two major contributions. The first contribution is an extension of the SPDE approach using bivariate splines (Lai and Schumaker, 2007) to allow more efficient and flexible treatment to the latent field. The second contribution is an application of the SPDE approach to combine multiple spatial surveys to achieve more accurate inference and spatial prediction. The second problem of uncertainty propagation is tackled using statistical emulation. The primary challenge here is the high dimensionality in the input space. We propose a joint framework for the high-dimensional emulation with a dimension reduction technique to overcome this hurdle. Though the thesis is motivated by, and focused on, tsunami research, most of the work is described within a more general context of spatial modelling and high-dimensional emulation and can be easily adapted to many other applications.

1.2 Thesis outline

In Chapter 2, we introduce an extension to the SPDE approach, which is also the technique employed in subsequent chapters to model the bathymetry. The SPDE approach is applicable in a wide range of problems based on the latent Gaussian models. It has been shown that computational efficiency can be gained by doing the computations using GMRFs as GFs can be seen as weak solutions to the corresponding SPDEs using piecewise linear finite elements. We introduce a new class of representations of GFs with bivariate splines instead of finite elements. This allows an easier implementation of piecewise polynomial representations of various degrees. It leads to GMRFs that can be inferred efficiently and can be easily extended to non-stationary fields. The solutions approximated with higher order bivariate splines converge faster, hence the computational burden can be alleviated. Numerical simulations using both real and simulated data also demonstrate that our approach provides more flexibility and efficiency when dealing with large scale and complicated spatial data. This chapter is based on the published work Liu et al.

(2015) jointly with S. Guillas and M.-J. Lai.

In Chapter 3, we deal with the statistical emulation for computer models with high-dimensional inputs. High accuracy complex computer models usually require large resources in time and memory to produce realistic results. Statistical emulators are computationally cheap approximations of such simulators. They are built to replace simulators for various purposes, such as the propagation of uncertainties from inputs to outputs or the calibration of some internal parameters against observations. However, when the input space is of high dimension, the construction of an emulator can become prohibitively expensive. We introduce a joint framework merging emulation with dimension reduction in order to overcome this hurdle. The gradient-based kernel dimension reduction technique is chosen due to its ability to extract drastically lower dimensions with little loss in information and its wide capability in various problems without any strong assumptions on the distribution and variable types. The Gaussian process emulation technique is combined with this dimension reduction approach. Our proposed framework therefore provides an answer to the dimension reduction issue in emulation for a wide range of problems that cannot be tackled at the moment. Theoretical properties of the approximation are explored. We demonstrate the efficiency, accuracy and advantages over other methods of the proposed approach on an elliptic PDE. We also present a realistic application to tsunami modelling. The uncertainties in the bathymetry are modelled as high-dimensional realisations of a spatial process using the SPDE approach. Our dimension-reduced emulation enables us to compute the impact of these uncertainties on resulting possible tsunami wave heights near-shore and on-shore. Considering an uncertain earthquake source, we observe a significant increase in the spread of uncertainties in the tsunami heights due to the contribution of the bathymetry uncertainties to the overall uncertainty budget. These results highlight the need to reduce uncertainties in the bathymetry in early warnings and hazard assessments. This chapter is based on the submitted work Liu and Guillas (2016).

In Chapter 4, we consider the geostatistical inference for multiple spatial surveys with a primary focus on the bathymetric surveys. Data from various surveys

are usually merged to construct bathymetric data products. These surveys may differ in many aspects including survey technique, coverage, accuracy and resolution. It is not advisable to just combine various data sources together and fit one single model regardless of the differentiation. We propose a joint hierarchical model based on the SPDE approach. This model allows us to make inference for the common underlying spatial process combining all the information from various surveys, with the flexibility to model each or a group of similar surveys separately in order to account for the respective characteristics. The proposed model also includes the preferential sampling feature when the sampling locations are stochastically dependent on the underlying spatial process. We illustrate the proposed method on simulated Gaussian fields and show that it makes the inference more accurate by considering the different features in multiple surveys. The joint model is also applied into the geostatistical mapping with a realistic bathymetry data set.

Chapter 5 consists of some conclusive discussion and future work. Appendix A includes theoretical proofs of the results in Chapter 2. The work in this thesis is closely related to the tsunami hazard assessment and the associated possible financial losses using catastrophe models. Some discussion and related projects are also included. Appendix B contains some work in the proof-of-concept study of tsunami risk for Cascadia region including data acquisition of bathymetry and topography, mesh generation and initial numerical simulations. The work in Appendix C is conducted for the commercial evaluation of the research findings about the impact of the uncertainties in the bathymetry on tsunami hazard, funded by the UCL Advances Enterprise Scholarship and EPSRC D2U project. The main focus is on building accurate and reliable tsunami hazard model for Catastrophe modelling on the Oasis LMF platform. This part includes some technical test of the Oasis platform as well as illustration of the potential financial consequences of the uncertainties in the bathymetry.

Chapter 2

Efficient Spatial Modelling Using the SPDE Approach with Bivariate Splines

2.1 Introduction

2.1.1 Latent Gaussian model with INLA

Gaussian fields (GFs) are at the core of spatial statistics, especially in the class of structured additive regression models, named latent Gaussian models, which are flexible and extensively used (Banerjee et al., 2004, Cressie, 1993, Diggle and Ribeiro, 2007). Suppose we have response variables y_i at locations $\mathbf{s}_i$, $i = 1, ..., n$, the hierarchical latent Gaussian model can be written as

$$\begin{aligned} y_i|x_i, \boldsymbol{\theta} &\sim P(y_i|x_i, \boldsymbol{\theta}) \\ \mathbf{x} &\sim \mathrm{N}(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})) \\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}) \end{aligned} \tag{2.1}$$

where $P(\cdot)$ is the conditional distribution of y_i given x_i and $\boldsymbol{\theta}$ which is the vector of all parameters relating the model, $\mathbf{x} = (x_1, ..., x_n)'$ is an unobserved spatial process, the latent Gaussian field, with mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, and $\pi(\cdot)$ is a prior of the parameters $\boldsymbol{\theta}$. The hierarchical structure provides more flexibility to describe the data features. At the same time, the underlying latent field is assumed to be

Gaussian with some good established properties and computational convenience. However, when making statistical inference, it is usually needed to evaluate the probability density function of the latent Gaussian field

$$\pi(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}.$$

The computations on dense matrices, e.g. the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, are typically of order $\mathcal{O}(n^3)$. Rue et al. (2009) overcome this computational hurdle by applying several innovations. They approximate Bayesian inference in latent Gaussian models by assuming that the latent field is Gaussian Markov random field (GMRF). The efficiency comes from the sparse structure in the precision matrix of a GMRF. Suppose $\mathbf{x} = (x_1, ..., x_n)'$ is a discontinuous indexed GMRF, then its precision matrix $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$ is sparse since for $i \neq j$, $\mathbf{Q}_{ij} \neq 0$ if and only if x_i and x_j are neighbours (Rue and Held, 2004). The sparse structure of $\mathbf{Q}$ makes its Cholesky decomposition, $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$ where $\mathbf{T}$ is a lower triangular matrix, more efficient. The sparsity of $\mathbf{Q}$ can be passed into $\mathbf{L}$ and only non-zero elements need to be computed. In most cases, the sparsity could be increased further by reordering $\mathbf{x}$ properly. The computational cost of the Cholesky decomposition is typically $\mathcal{O}(n)$ for one dimensional GMRF, $\mathcal{O}(n^{3/2})$ for two dimensions and $\mathcal{O}(n^2)$ for three dimensions. The fast Cholesky decomposition can be used to speed up the inference in many aspects. For example, the components in the likelihood can be calculated easily, e.g. $\log|\boldsymbol{\Sigma}| = -\log|\mathbf{Q}| = -2\sum_{i=1}^{n}\log L_{ii}$. It is also more efficient to draw samples from a GMRF with precision matrix $\mathbf{Q}$ by just solving the linear system $\mathbf{L}^T\mathbf{x} = \mathbf{z}$ for some random variables $\mathbf{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_n)$. Taking advantage of the sparsity of GMRFs, Rue et al. (2009) proposed the integrated nested Laplace approximations (INLA) approach which produces faster inference than simulation based approaches such as Markov chain Monte Carlo (MCMC) methods. Considering the latent Gaussian model (2.1) where the latent field is GMRF and the number of hyperparameters is small, the INLA approach approximates full Bayesian inference by directly calculate the approximate posterior distributions. The posterior

distribution of the unknowns is

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

The aim is to approximate the posterior marginals $\pi(x_i|\mathbf{y})$, $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $\pi(\theta_j|\mathbf{y})$. The INLA method is based on the Laplace approximation of the posterior for $\boldsymbol{\theta}|\mathbf{y}$,

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \approx \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})},$$

where $\mathbf{x}^*(\boldsymbol{\theta}) = \text{argmax}_{\mathbf{x}}\, \pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$. The posterior for the latent Gaussian field is then approximated with a Gaussian approximation so that

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} \approx \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})}\Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})},$$

where $\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ is the Gaussian approximation to $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ and the mode can be found by using numerical optimisation. The marginal posteriors for each component of $\boldsymbol{\theta}$ and $\mathbf{x}$ can be calculated using numerical integration over $\boldsymbol{\theta}$,

$$\pi(\theta_i|\mathbf{y}) \approx \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\, \boldsymbol{\theta}_{-i},$$
$$\pi(x_i|\mathbf{y}) \approx \int \tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y})\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})\, \mathrm{d}\, \boldsymbol{\theta}.$$

Therefore, the INLA approach performs direct numerical calculation of the posterior densities, avoiding time-consuming MCMC sampling. It has been illustrated to be efficient and effective for the latent Gaussian models. For more details about this approach, see Rue et al. (2009) and Martins et al. (2013).

2.1.2 SPDE approach

Though the latent Gaussian models based on GMRFs are computationally efficient, most of the GMRF models are too simple hence their applications are restricted. As discussed in Lindgren et al. (2011), it is difficult to parameterise the sparse precision matrix with presumed correlation between any two sites and the use of simple neighbourhood makes it unclear how wide the useful GMRF models family

is. Lindgren et al. (2011) constructed an explicit link between GFs and GMRFs. Then it is possible to model with GFs which are widely applicable in many applications while doing computations with GMRFs. They considered the GFs with Matérn covariance function,

$$r(\mathbf{u},\mathbf{v}) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa\|\mathbf{v}-\mathbf{u}\|)^{\nu}K_{\nu}(\kappa\|\mathbf{v}-\mathbf{u}\|), \tag{2.2}$$

where $\|\mathbf{v}-\mathbf{u}\|$ is the Euclidean distance between two locations $\mathbf{u}$ and $\mathbf{v} \in \mathbb{R}^D$, K_ν is the modified Bessel function of the second kind and order $\nu > 0$, $\kappa > 0$ controls the nominal correlation range through $\rho = \sqrt{8\nu}/\kappa$ corresponding to correlations near 0.1 at the Euclidean distance ρ, and σ^2 is the marginal variance. The integer value of ν determines the mean-square differentiability of the underlying process. Generally speaking, a Gaussian process with the Matérn covariance (2.2) has sample paths that are $\lfloor \nu - 1 \rfloor$ times differentiable (Paciorek and Schervish, 2004). The relationships between different parameters and the Matérn covariance function as well as some samples drawn from a one-dimensional Matérn field when $\sigma^2 = 1$ are illustrated in Figure 2.1.

Lindgren et al. (2011) noticed that a Gaussian field $x(\mathbf{u})$ with the Matérn covariance (2.2) is a solution to the linear fractional SPDE

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau x(\mathbf{u})) = W(\mathbf{u}), \quad \mathbf{u} \in \mathbb{R}^D, \ \ \alpha = \nu + D/2, \ \ \kappa > 0, \ \ \nu > 0, \tag{2.3}$$

where the innovation process W is spatial Gaussian white noise with unit variance (Whittle, 1954, 1963), $\Delta = \sum_{i=1}^{D} \frac{\partial^2}{\partial x_i^2}$ is the Laplacian operator, and τ controls the marginal variance through the relationship

$$\tau^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + D/2)(4\pi)^{D/2}\kappa^{2\nu}\sigma^2}.$$

Denoting the inner product of two functions f and g on $\mathbb{R}^D$ as $\langle f, g\rangle = \int_{\mathbb{R}^D} f(\mathbf{u})g(\mathbf{u})\,\mathrm{d}\,\mathbf{u}$, we consider the stochastic weak formulation of the SPDE (2.3)

$$\left\{\langle \phi_t, (\kappa^2 - \Delta)^{\alpha/2}\tau x\rangle, t = 1, ..., n_t\right\} \overset{\mathrm{d}}{=} \left\{\langle \phi_t, W\rangle, t = 1, ..., n_t\right\}, \tag{2.4}$$

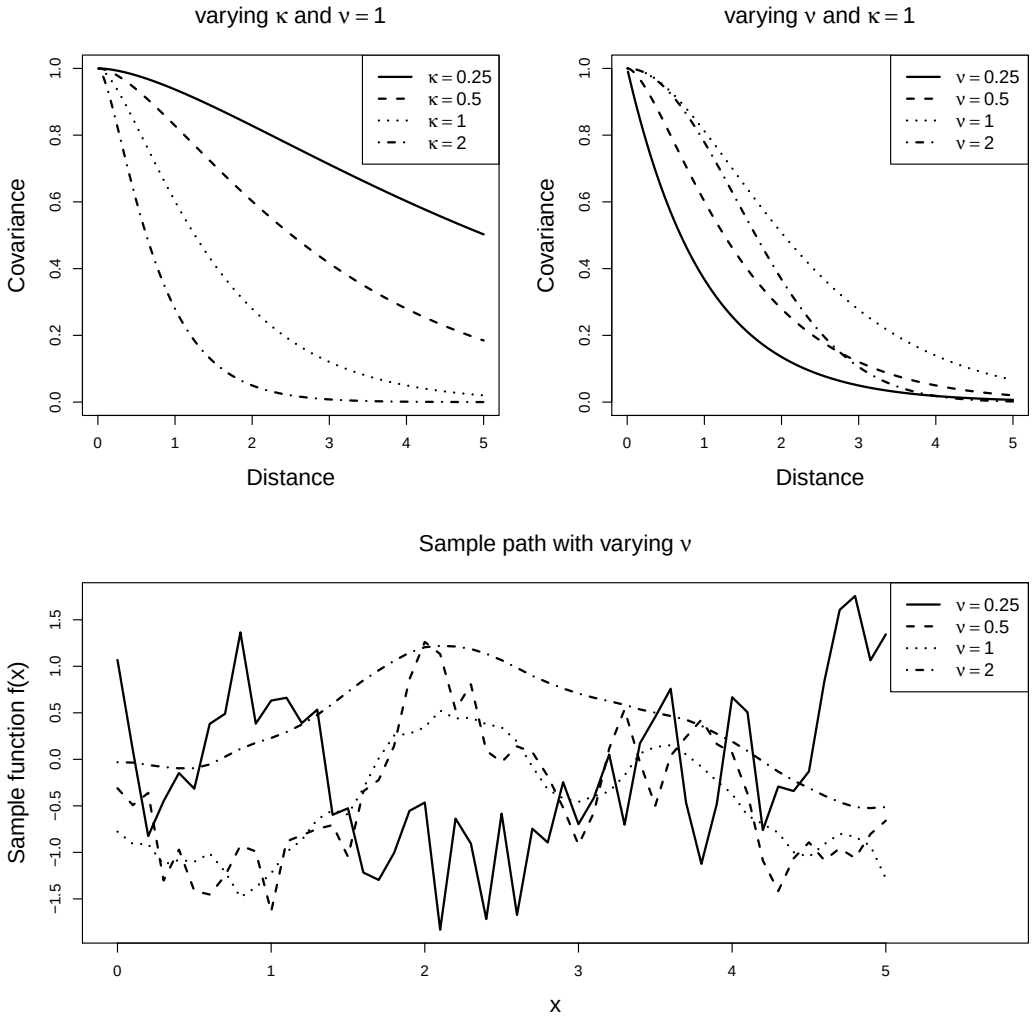


Figure 2.1: Matérn covariance functions with varying κ and fixed ν (top left), and varying ν and fixed κ (top right); random samples drawn from a Gaussian process with mean zero and Matérn covariance for different ν where $\kappa = 1$(bottom).

for every finite set of suitable test functions $\{\phi_t(\mathbf{u}), t = 1, ..., n_t\}$, where '$\stackrel{\mathrm{d}}{=}$' denotes equality in distribution (Walsh, 1986). Lindgren et al. (2011) constructed a finite element representation (Brenner and Scott, 2008) of the Gaussian random field over an unstructured triangulation of the form

$$x_h(\mathbf{u}) = \sum_{k=1}^{n} w_k \psi_k(\mathbf{u}), \tag{2.5}$$

where $\{\psi_k\}_{k=1}^n$ are piecewise linear basis functions. By requiring (2.4) to hold for only a specific set of test functions, they showed that the Gaussian weights $\{w_k\}_{k=1}^n$

are GMRFs when $\alpha = 1$, and can be approximated with GMRFs when $\alpha \geq 2$. Define the $n \times n$ matrices $\mathbf{C}$, $\mathbf{G}$ and $\mathbf{K}$ with (i, j)-th entry as,

$$\begin{aligned} \mathbf{C}_{ij} &= \langle \varphi_i, \varphi_j \rangle, \\ \mathbf{G}_{ij} &= \langle \nabla\varphi_i, \nabla\varphi_j \rangle, \\ (\mathbf{K})_{ij} &= \kappa^2 C_{ij} + G_{ij}, \end{aligned} \tag{2.6}$$

for $i, j = 1, 2, ..., n$. Using Neumann boundary conditions (zero normal derivative at the boundary, $\partial_{\mathbf{n}} x = 0$), Lindgren et al. (2011) have the result quoted below.

Result 1. *Let $\mathbf{Q}_\alpha$ be the precision matrix for the Gaussian weights $\mathbf{w} = \{w_k\}_{k=1}^n$ as defined in equation (2.5) for $\alpha = 1, 2, ...$, as a function of κ^2 and τ. Then*

$$\begin{aligned} \mathbf{Q}_1 &= \tau\mathbf{K}, \\ \mathbf{Q}_2 &= \tau^2\mathbf{K}\mathbf{C}^{-1}\mathbf{K}, \\ \mathbf{Q}_\alpha &= \tau^2\mathbf{K}\mathbf{C}^{-1}\mathbf{Q}_{\alpha-2}\mathbf{C}^{-1}\mathbf{K}, \quad \text{for } \alpha = 3, 4, ... \end{aligned} \tag{2.7}$$

Thus a Gaussian field $x(\mathbf{u})$ with Matérn covariance can be modelled as a linear predictor of $\mathbf{w}$ whose precision matrix has been shown in Result 1. The inference for a GF can be carried out on the associated GMRF and the computational efficiency can be improved dramatically, especially with the INLA approach. This work is closely related to the spatial spline regression models by Sangalli et al. (2013) where a spatial surface is approximated with finite elements as well. Another recent related work is Nychka et al. (2015), where the authors proposed a representation of a random field using multi-resolution radial basis functions on a regular grid. They also assumed that the coefficients associated with the basis functions to be distributed according to a GMRF to speed up the computation.

It is stated in Lindgren et al. (2011) and Simpson et al. (2012) that the convergence rate of a finite element approximation to the full solution to the SPDE (2.3) is of order $\mathcal{O}(h^2)$ where h is the length of longest edge in the triangulation. Hence the convergence can be achieved by refining the underlying triangulation which is usually called the h-version finite elements. An alternative is to increase the ap-

proximation order over any fixed triangulation with higher degree polynomials over each triangle, which is called the p-version finite elements (the degree of polynomials is usually denoted by p). It has been illustrated that the convergence rate of the p-version cannot be worse than the h-version in most cases (Babuska et al., 1981). To do so, multivariate splines over triangulations can be employed instead of conventional finite elements. This provides a flexible and easy construction of splines with piecewise polynomials of various degrees and smoothness. Basic concepts and theories of multivariate splines can be found in the monograph by Lai and Schumaker (2007). Multivariate splines have been shown to be more efficient and flexible than conventional finite element method in data fitting problems and solving PDEs, see Awanou et al. (2006). It has been applied in spatial statistics. For example, Guillas and Lai (2010) introduced a spatial data analysis model with bivariate splines by penalising the roughness with a partial differential operator; this has been demonstrated to be more efficient and accurate than thin-plate splines (Wood, 2003) in the application of ozone concentration forecasting (Ettinger et al., 2012). In this paper, we introduce bivariate splines to represent the GFs on $\mathbb{R}^2$ and show its advantages over the piecewise linear finite elements in Lindgren et al. (2011). Within our framework of the SPDE approach using bivariate splines, it is allowed to choose piecewise polynomial representations of arbitrary degrees to adapt to the various data structures and features. It also makes the inference more computationally efficient.

The rest of this chapter is structured as follows. In Section 2.2, some basics of bivariate splines in the Bernstein form (B-form) are reviewed first. Then we show how to link the GFs with GMRFs within the framework of bivariate splines, establish the theoretical properties of the bivariate spline approximations and discuss extensions to non-stationary fields. In Section 2.3, we conduct several numerical simulations to illustrate our method and compare with the approach of Lindgren et al. (2011) on both real and simulated data sets. Section 2.4 consists of conclusion and discussion. Proofs for some results are in Appendix A.

2.2 SPDE approach using bivariate splines

2.2.1 B-form bivariate splines

Triangulation is commonly used in finite elements and multivariate splines to construct functional representations of a complicated surface over complex and irregular domain. It is a set of triangles $\mathbf{\Delta} = \{T_1, ..., T_N\}$ such that if any two triangles in $\mathbf{\Delta}$ intersect, then the intersection must be either a common vertex or a common edge. Figure 2.2 presents two sets of triangles that discretise the same region. The one in the left forms a triangulation but the other in the right does not form a triangulation since the intersection between the upper triangle and each of the two lower triangles is only part of its edge.

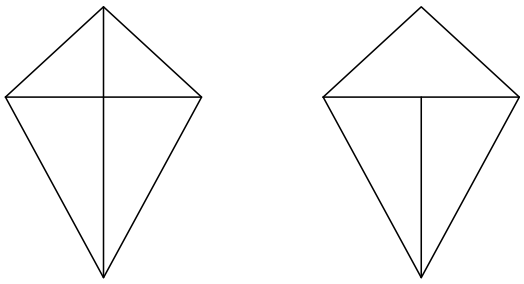


Figure 2.2: An example of a triangulation (left) and a set of triangles (right) that do not form a triangulation.

Let $\mathbf{\Delta}$ be a triangulation of a bounded domain $\Omega \subset \mathbb{R}^2$. We consider the continuous spline spaces

$$S_d^0(\mathbf{\Delta}) = \{s \in C^0(\Omega), s|_T \in \mathcal{P}_d, \forall T \in \mathbf{\Delta}\},$$

where $\mathcal{P}_d$ is the space of bivariate polynomials of degree $d \geq 1$, $C^0(\Omega)$ is the space of all continuous functions on Ω. For any $d \geq 1$, the spline space $S_d^0(\mathbf{\Delta})$ contains all possible continuous spline functions which are bivariate polynomials of degree d over each triangle $T \in \mathbf{\Delta}$. The B-form representation of splines in $S_d^0(\mathbf{\Delta})$ proposed by Awanou et al. (2006) is employed. We only give a brief introduction to the bivariate splines here. For more complete and in-depth explanations, see Lai and

Schumaker (2007).

Let $T = \langle \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \rangle$ be a non-degenerate (i.e. with non-zero area) triangle with vertices $\mathbf{v}_1 = (x_1, y_1)$, $\mathbf{v}_2 = (x_2, y_2)$ and $\mathbf{v}_3 = (x_3, y_3)$. Then every point $\mathbf{v} = (x, y) \in \mathbb{R}^2$ has a unique representation in the form

$$\mathbf{v} = b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + b_3\mathbf{v}_3, \tag{2.8}$$

with $b_1 + b_2 + b_3 = 1$, where b_1, b_2 and b_3 are named the barycentric coordinates of the point $\mathbf{v} = (x, y)$ relative to the triangle T. The polynomials

$$B_{ijk}^{T,d}(\mathbf{v}) = \frac{d}{i!j!k!} b_1^i b_2^j b_3^k, \; i + j + k = d, \tag{2.9}$$

are called the Bernstein polynomials of degree d relative to triangle T. We may denote it as B_{ijk}^d for simplicity when the underlying triangle T is referred as general or clear in the context.

To evaluate a polynomial of degree d in B-form over any triangle, say $p = \sum_{i+j+k=d} c_{ijk} B_{ijk}^d$, at the point $\mathbf{v} = (x, y)$ whose barycentric coordinates are $b = (b_1, b_2, b_3)$ with $b_1 + b_2 + b_3 = 1$, let $c_{ijk}^{(0)} = c_{ijk}$ and for all $l = 1, ..., d$,

$$c_{ijk}^{(l)} = b_1 c_{i+1,j,k}^{(l-1)} + b_2 c_{i,j+1,k}^{(l-1)} + b_3 c_{i,j,k+1}^{(l-1)}.$$

For $i + j + k = d - l$, we have

$$p(\mathbf{v}) = \sum_{i+j+k=d-l} c_{ijk}^{(l)} B_{ijk}^{d-l}(\mathbf{v}),$$

for all $0 \leq l \leq d$. In particular, $p(\mathbf{v}) = c_{000}^{(d)}$. This is called the de Casteljau algorithm (Lai and Schumaker, 2007).

Each vector $\mathbf{u}$ can be uniquely described by a triple (a_1, a_2, a_3) called directional coordinates of $\mathbf{u}$, that is $a_i = \alpha_i - \beta_i$, $i = 1, 2, 3$, where $(\alpha_1, \alpha_2, \alpha_3)$ and $(\beta_1, \beta_2, \beta_3)$ are the barycentric coordinates of two points $\boldsymbol{\omega}$ and $\tilde{\boldsymbol{\omega}}$ such that $\mathbf{u} = \boldsymbol{\omega} - \tilde{\boldsymbol{\omega}}$. It is easy to see that the barycentric coordinates of a point sum to 1,

while the directional coordinates of a vector sum to 0. Suppose $\mathbf{u}$ is a vector in $\mathbb{R}^2$ whose directional coordinates are $a = (a_1, a_2, a_3)$, then for $i+j+k = d$, we define the directional derivative of B^d_{ijk} at location $\mathbf{v}$ with respect to directional vector $\mathbf{u}$ to be

$$D_u B^d_{ijk}(\mathbf{v}) = d\left[a_1 B^{d-1}_{i-1,j,k}(\mathbf{v}) + a_2 B^{d-1}_{i,j-1,k}(\mathbf{v}) + a_3 B^{d-1}_{i,j,k-1}(\mathbf{v})\right]. \tag{2.10}$$

The integrals and inner products of the Bernstein polynomials can be calculated precisely as presented in the following lemma.

Lemma 1. *Let* $p = \sum_{i+j+k=d} c_{ijk} B^d_{ijk}$ *be a polynomial of degree* d *on triangle* T *(with area* A_T*), then*

$$\int_T p(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y = \frac{A_T}{\binom{d+2}{2}} \sum_{i+j+k=d} c_{ijk}. \tag{2.11}$$

Let $q = \sum_{\nu+\mu+\kappa=d} \tilde{c}_{\nu\mu\kappa} B^d_{\nu\mu\kappa}$ *be another polynomial of degree* d *on triangle* T*, then the inner product of* p *and* q *is*

$$\int_T p(x,y)q(x,y)\,\mathrm{d}\,x\,\mathrm{d}\,y = \frac{A_T}{\binom{2d}{d}\binom{2d+2}{2}} \sum_{\substack{i+j+k=d \\ \nu+\mu+\kappa=d}} \binom{i+\nu}{i}\binom{j+\mu}{j}\binom{k+\kappa}{k} c_{ijk}\tilde{c}_{\nu\mu\kappa}. \tag{2.12}$$

For each spline function $s \in S^0_d(\boldsymbol{\Delta})$, we can write

$$s|_T = \sum_{i+j+k=d} c^T_{ijk} B^{T,d}_{ijk}, \quad T \in \boldsymbol{\Delta},$$

where the coefficients $\mathbf{c} = \{c^T_{ijk}, i+j+k = d, T \in \boldsymbol{\Delta}\}$ are called B-coefficients of s. Note that linear finite elements are typical splines in $S^0_1(\boldsymbol{\Delta})$. For the spline space $S^0_d(\boldsymbol{\Delta})$, the domain points are defined to be the set

$$\mathcal{D}_{d,\boldsymbol{\Delta}} = \{\xi_{ijk} = (i\mathbf{v}_1 + j\mathbf{v}_2 + k\mathbf{v}_3)/d, i+j+k = d, T = \langle \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\rangle \in \boldsymbol{\Delta}\}.$$

Therefore the spline function can also be denoted by

$$s|_T = \sum_{\xi \in \mathcal{D}_{d,T}} c_\xi B_\xi^{T,d},$$

where $B_\xi^{T,d}$ stands for $B_{ijk}^{T,d}$ for $\xi = \xi_{ijk} \in \mathcal{D}_{d,T}$ and c_ξ is the corresponding B-coefficient c_{ijk}. Note that since s is continuous, if ξ lies on an edge shared by two different triangles T and $\tilde{T}$, then the corresponding coefficients c_ξ for $s|_T$ and $s|_{\tilde{T}}$ should be the same. Then we show that the basis for $S_d^0(\mathbf{\Delta})$ can be constructed easily with spline functions in $S_d^0(\mathbf{\Delta})$ with specific B-coefficients. For each $\xi \in \mathcal{D}_{d,\mathbf{\Delta}}$, let ψ_ξ be the spline in $S_d^0(\mathbf{\Delta})$ having all zero B-coefficients except for $c_\xi = 1$, then we have the following result

Lemma 2. *The set of splines $\mathcal{B} = \{\psi_\xi, \xi \in \mathcal{D}_{d,\mathbf{\Delta}}\}$ forms a basis for the spline space $S_d^0(\mathbf{\Delta})$ which satisfies $\psi_\xi(\mathbf{v}) \geq 0$ and $\sum_{\xi \in \mathcal{D}_{d,\mathbf{\Delta}}} \psi_\xi(\mathbf{v}) = 1$ for all $\mathbf{v} \in \Omega$.*

It is obvious that ψ_ξ is identically zero on all triangles that do not contain ξ since the corresponding B-coefficients are all zeros so that ψ_ξ is locally supported.

2.2.2 SPDE modelling with B-form bivariate splines

Let $\{\psi_1, \psi_2, ..., \psi_m\}$ be a set of locally supported basis functions of $S_d^0(\mathbf{\Delta})$ for any $d \geq 1$, where $m = \dim S_d^0(\mathbf{\Delta})$, as stated in Lemma 2, on a triangulation of a bounded domain $\Omega \subset \mathbb{R}^2$. For any $h = 1, ..., m$, the corresponding B-coefficients of ψ_h are denoted by $\mathbf{c}_h$. The results in this chapter hold for any triangulations. But the quality of approximations depend on the triangulation properties. In practice, we suggest the Delaunay triangulations that are chosen to maximize the minimum interior triangle angle following Lindgren et al. (2011).

Then we can construct a bivariate spline representation of the solution to SPDE (2.3) in the spline space $S_d^0(\mathbf{\Delta})$ as

$$x_{\mathbf{\Delta}}(\mathbf{u}) = \sum_{h=1}^{m} w_h \psi_h(\mathbf{u}). \tag{2.13}$$

Following Lindgren et al. (2011) we approximate a weak solution to the SPDE with respect to the spline space $S_d^0(\mathbf{\Delta})$ by finding the distribution of the weights

$\{w_h, h = 1, ..., m\}$ that fulfils the stochastic weak formulation (2.4) for only a specific set of test functions such that the integrals at both sides of (2.4) exist. The distribution of the approximate solution $x_{\boldsymbol{\Delta}}(\mathbf{u})$ can be obtained through the stochastic weights. Specifically, we choose $\phi_h = (\kappa^2 - \Delta)^{1/2}\psi_h$ for $\alpha = 1$ leading to the least squares solution. For $\alpha = 2$, we can choose either $\phi_h = \psi_h$ for any $d \geq 1$ or $\phi_h = (\kappa^2 - \Delta)\psi_h$ for $d \geq 2$, leading to the Galerkin or least squares solution respectively. For $\alpha \geq 3$, if we let $\alpha = 2$ on the left-hand side of the SPDE (2.3) then the right-hand side is a Gaussian process generated by the operator $(\kappa^2 - \Delta)^{(\alpha-2)/2}$. Then we can choose $\phi_h = \psi_h$ for this innovative SPDE. Hence we get a recursive Galerkin solutions ending with either $\alpha = 1$ or 2. We also assume appropriate boundary conditions to avoid the solutions in the null space of the differential operator. Throughout this chapter, the Neumann condition (zero normal derivative at the boundary) is imposed. Then we have the main results as below.

Theorem 1. *The vector of weights* $\mathbf{w} = (w_1, ..., w_m)^T$ *of bivariate spline representation (2.13) is Gaussian with mean zero and the precision matrix* $\mathbf{Q}_\alpha$ *that are given as follows:*

(1) for $\alpha = 1$*,*

$$\mathbf{Q}_1 = \tau^2(\kappa^2\mathbf{M} + \mathbf{K}),$$

(2) for $\alpha = 2$*,*

$$\mathbf{Q}_2^G = \tau^2(\kappa^4\mathbf{M} + 2\kappa^2\mathbf{K} + \mathbf{K}\mathbf{M}^{-1}\mathbf{K}),$$

$$\mathbf{Q}_2^{LS} = \tau^2(\kappa^4\mathbf{M} + 2\kappa^2\mathbf{K} + \mathbf{R}),$$

where $\mathbf{Q}_2^G$ *and* $\mathbf{Q}_2^{LS}$ *are the Galerkin and least squares solutions respectively,*

(3) for $\alpha \geq 3$*,*

$$\mathbf{Q}_\alpha = \kappa^4\mathbf{Q}_{\alpha-2} + \kappa^2(\mathbf{Q}_{\alpha-2}\mathbf{M}^{-1}\mathbf{K} + \mathbf{K}\mathbf{M}^{-1}\mathbf{Q}_{\alpha-2}) + \mathbf{K}\mathbf{M}^{-1}\mathbf{Q}_{\alpha-2}\mathbf{M}^{-1}\mathbf{K},$$

where

$$\mathbf{M} = \mathbf{C}'\mathbf{M_0}\mathbf{C}, \;\; \mathbf{K} = \mathbf{C}'\mathbf{K_0}\mathbf{C}, \;\; \mathbf{R} = \mathbf{C}'\mathbf{R_0}\mathbf{C},$$

and $\mathbf{M_0} = \mathrm{diag}(\mathbf{M}_T, T \in \boldsymbol{\Delta})$, $\mathbf{K_0} = \mathrm{diag}(\mathbf{K}_T, T \in \boldsymbol{\Delta})$ *and* $\mathbf{R_0} = \mathrm{diag}(\mathbf{R}_T, T \in \boldsymbol{\Delta})$ *are block diagonal square matrix with square blocks*

$$\mathbf{M}_T = \left[\int_T B^T_{ijk}(x,y) B^T_{i'j'k'}(x,y) \,\mathrm{d}\, x \,\mathrm{d}\, y \right]^{i'+j'+k'=d}_{i+j+k=d},$$

$$\mathbf{K}_T = \left[\int_T \nabla B^T_{ijk}(x,y) \nabla B^T_{i'j'k'}(x,y) \,\mathrm{d}\, x \,\mathrm{d}\, y \right]^{i'+j'+k'=d}_{i+j+k=d},$$

and

$$\mathbf{R}_T = \left[\int_T \Delta B^T_{ijk}(x,y) \Delta B^T_{i'j'k'}(x,y) \,\mathrm{d}\, x \,\mathrm{d}\, y \right]^{i'+j'+k'=d}_{i+j+k=d},$$

respectively and $\mathbf{C} = (\mathbf{c}_1, ..., \mathbf{c}_m)$ *whose* h*-th column is the B-coefficient vector of basis function* ψ_h*.*

Since the basis functions $\{\psi_1, \psi_2, ..., \psi_m\}$ are locally supported in $S^0_d(\boldsymbol{\Delta})$, the matrices $\mathbf{M}$, $\mathbf{K}$ and $\mathbf{R}$ are guaranteed to be sparse. However in the Galerkin solution for $\alpha = 2$ and recursive solutions for $\alpha \geq 3$, the inverse matrix $\mathbf{M}^{-1}$ is not necessarily sparse, making the precision matrix dense. The mass lumping technique (Chen and Thomée, 1985) can be applied by replacing $\mathbf{M}$ with a diagonal matrix $\tilde{\mathbf{M}}$ whose elements are the sum of each row of $\mathbf{M}$, i.e. $\tilde{\mathbf{M}}_{ii} = \sum_j \mathbf{M}_{ij}$. The same technique is also deployed by Lindgren et al. (2011) and they discussed the implications in detail in their Appendix C.5 showing that the convergence rate would not be affected. We will discuss the properties of the mass lumping approximation in our case later. Therefore, the precision matrix is sparse and the underlying coefficients $\mathbf{w}$ are approximated with a GMRF.

2.2.3 Approximation properties

Define the Hilbert space H^1 associated with the differential operator $(\kappa^2 - \Delta)$ to be the space of square integrable functions $f(x,y)$ for which $\|f\|^2_{H^1} = \kappa^2 \int_\Omega f(x,y)^2 \,\mathrm{d}\, x \,\mathrm{d}\, y + \int_\Omega \nabla f(x,y) \cdot \nabla f(x,y) \,\mathrm{d}\, x \,\mathrm{d}\, y$ is finite following Lindgren et al. (2011). Approximation results for bivariate splines, e.g. Th. 5.19 in Lai and Schumaker (2007), show that the bivariate spline space $S^0_d(\boldsymbol{\Delta})$ for any $d \geq 1$ spanned by a finite set of basis functions $\{\psi_1, ..., \psi_m\}$ is dense in H^1: for every

$f \in H^1$, there is a sequence $\{f_m\}$, $f_m \in S_d^0(\mathbf{\Delta})$ such that $\lim_{m\to\infty} \|f - f_m\|_{H^1} = 0$ where the limit scenario $m \to \infty$ corresponds to $|\mathbf{\Delta}| \to 0$ where $|\mathbf{\Delta}|$ is the length of the longest edge in the triangulation $\mathbf{\Delta}$. Using this fact, it follows directly from the Th. 3-4 in Appendix C.2 of Lindgren et al. (2011) that, the bivariate spline approximation $x_{\mathbf{\Delta}}$ converges weakly to the weak solution to the SPDE. Note that the weak convergence of $x_{\mathbf{\Delta}}$ obtained for $\mathbf{Q}_2^{LS}$ cannot be derived directly but can be easily proved in the same fashion with just a few modifications. In addition, we can derive rates of convergence results. We first define the associated Sobolev space on Ω in $\mathbb{R}^2$ for any $1 \le q \le \infty$ and $d \ge 1$ as

$$W_q^d(\Omega) = \{f : \|f\|_{d,q,\Omega} < \infty\},$$

where

$$\|f\|_{d,q,\Omega} = \begin{cases} \left(\sum_{k=0}^{d} |f|_{k,q,\Omega}^q\right)^{1/q}, & 1 \le q < \infty \\ \sum_{k=0}^{d} |f|_{k,\infty,\Omega}, & q = \infty, \end{cases}$$

with

$$|f|_{k,q,\Omega} = \begin{cases} \left(\sum_{\nu+\mu=k} \|D_x^\nu D_y^\mu f\|_{q,\Omega}^q\right)^{1/q}, & 1 \le q < \infty \\ \max_{\nu+\mu=k} \|D_x^\nu D_y^\mu f\|_{\infty,\Omega}, & q = \infty, \end{cases}$$

and

$$\|f\|_{q,\Omega} = \begin{cases} \left(\int_\Omega |f(u)|^q du\right)^{1/q}, & 1 \le q < \infty, \\ \operatorname{ess\,sup}_{u\in\Omega} |f(u)|, & q = \infty. \end{cases}$$

Then we have the proposition below regarding to the Galerkin solutions when $\alpha = 2$.

Proposition 1. *Let $L = (\kappa^2 - \Delta)$, $x_{\mathbf{\Delta}}(\mathbf{u})$ is the bivariate spline approximation of the random Gaussian field $x(\mathbf{u})$ in the spline space $S_d^0(\mathbf{\Delta})$, $d \ge 1$. Then for any $f \in H^1 \cap W_2^{m+1}(\Omega)$ with $1 \le m \le d$, we have*

$$\mathrm{E}\left(\int_\Omega f(\mathbf{u}) L(x(\mathbf{u}) - x_{\mathbf{\Delta}}(\mathbf{u}))\,\mathrm{d}\,\mathbf{u}\right)^2 \le K|\mathbf{\Delta}|^{m+1}|f|_{m+1,2,\Omega},$$

where K is a constant, $|\boldsymbol{\Delta}|$ is the length of the longest triangle edge in the triangulation $\boldsymbol{\Delta}$.

It is clear that we are able to achieve a faster convergence rate by using bivariate splines with higher degree d. For example, when $d = 3$ the convergence rate can be as high as $\mathcal{O}(|\boldsymbol{\Delta}|^4)$, which is two magnitude higher than $\mathcal{O}(|\boldsymbol{\Delta}|^2)$ in Lindgren et al. (2011).

As we have mentioned, the matrix $\mathbf{M}$ in the Galerkin solutions is lumped by replacing $\mathbf{M}$ with a diagonal matrix $\tilde{\mathbf{M}}$ which yields a Markov approximation $\tilde{x}_{\boldsymbol{\Delta}}$ to the bivariate spline solution $x_{\boldsymbol{\Delta}}$. Let f and g be test functions in H^1 and let $f_{\boldsymbol{\Delta}}$ and $g_{\boldsymbol{\Delta}}$ be their projections onto the bivariate spline space $S_d^0(\boldsymbol{\Delta})$ for any $d \geq 1$, with basis weights $\mathbf{w}_f$ and $\mathbf{w}_g$. Since the recursive algorithm for $\alpha \geq 3$ is based on $\alpha = 2$ at each iteration, here we only investigate the effect of the Markov approximation on the Galerkin solutions for $\alpha = 2$. When $\alpha = 2$, the difference between the covariances for the Markov approximation $\tilde{x}_{\boldsymbol{\Delta}}$ and the bivariate spline solution $x_{\boldsymbol{\Delta}}$ is

$$\begin{aligned}\epsilon_{\boldsymbol{\Delta}}(f_{\boldsymbol{\Delta}}, g_{\boldsymbol{\Delta}}) &= \mathrm{Cov}(\langle f, L\tilde{x}_{\boldsymbol{\Delta}}\rangle_\Omega, \langle g, L\tilde{x}_{\boldsymbol{\Delta}}\rangle_\Omega) - \mathrm{Cov}(\langle f, Lx_{\boldsymbol{\Delta}}\rangle_\Omega, \langle g, Lx_{\boldsymbol{\Delta}}\rangle_\Omega) \\ &= \mathbf{w}_f'\tilde{\mathbf{M}}\mathbf{w}_g - \mathbf{w}_f'\mathbf{M}\mathbf{w}_g.\end{aligned}$$

We have the following result showing that such a difference can be bounded.

Proposition 2. *For $f_{\boldsymbol{\Delta}}$, $g_{\boldsymbol{\Delta}} \in S_d^0(\boldsymbol{\Delta})$, we have*

$$|\epsilon_{\boldsymbol{\Delta}}(f_{\boldsymbol{\Delta}}, g_{\boldsymbol{\Delta}})| \leq K|\boldsymbol{\Delta}|^2$$

where K is a positive constant dependent on $\|f_{\boldsymbol{\Delta}}\|_{2,2,\Omega}$, $\|g_{\boldsymbol{\Delta}}\|_{2,2,\Omega}$, $\|f_{\boldsymbol{\Delta}}\|_{\infty,\Omega}$, $\|g_{\boldsymbol{\Delta}}\|_{\infty,\Omega}$ and $|\boldsymbol{\Delta}|$ is the length of the longest triangle edge in the triangulation $\boldsymbol{\Delta}$.

We can see that the mass lumping error is at most of order $\mathcal{O}(|\boldsymbol{\Delta}|^2)$. We are not sure whether this is the lowest bound in theory. Thus the overall approximation error of bivariate splines representations to the SPDE solutions is at most $\mathcal{O}(|\boldsymbol{\Delta}|^2)$, which is the same as Lindgren et al. (2011).

These results broadly reach an agreement with the numerical simulations in Bolin and Lindgren (2013). The authors have shown that the higher order splines are more efficient in covariance approximation but become less efficient using mass lumping approximation. It seems to suggest higher order bivariate splines may be less helpful to the convergence. However for practical applications, we want to make some points clear here. Firstly, the actual approximation errors in both propositions also depend on the unknown constants K. Secondly, as stated in Bolin and Lindgren (2013), parameter inference is also very important to the accuracy and efficiency in application, which could be affected by using different representations. Last but not least, higher order splines may lose efficiency in approximating the true Matérn field with mass lumping, but not necessarily in approximating the spatial field. Thus efficiency can still be improved in practice by using higher order bivariate splines. The numerical simulations later illustrate that higher order bivariate splines are more efficient in spatial prediction.

2.2.4 Non-stationary fields

Lindgren et al. (2011) showed that the SPDE (2.3) can be extended to a non-stationary version

$$(\kappa^2(\mathbf{u}) - \Delta)^{\alpha/2}(\tau(\mathbf{u})x(\mathbf{u})) = W(\mathbf{u}), \tag{2.14}$$

where the parameters κ^2 and τ are not constants but depend on the location $\mathbf{u}$. The two parameters are assumed to vary slowly and have the general form of low dimensional representations

$$\log(\kappa^2(\mathbf{u})) = \sum_{j=1}^{n_{\kappa^2}} \theta_j^{(\kappa^2)} B_j^{(\kappa^2)}(\mathbf{u}), \quad \log(\tau(\mathbf{u})) = \sum_{j=1}^{n_\tau} \theta_j^{(\tau)} B_j^{(\tau)}(\mathbf{u}),$$

where the number of smooth basis functions n_{κ^2} and n_τ should not be large to guarantee computational efficiency. The inner product can be approximated with $\langle \psi_t, \kappa^2 \psi_s \rangle \approx \kappa^2(\mathbf{u}_s^\star)\langle \psi_t, \psi_s \rangle$, where $\mathbf{u}_s^\star$ is some point in the support of ψ_s which can be chosen to be the domain point associated with the non-zero B-coefficients of

ψ_s. Defining the diagonal matrices

$$\boldsymbol{\kappa}^2 = \text{diag}(\kappa^2(\xi_h), h = 1, ..., m), \quad \boldsymbol{\tau} = \text{diag}(\tau(\xi_h), h = 1, ..., m),$$

where ξ_h is the domain point associated with the non-zero B-coefficients of basis function ψ_h for $h = 1, ..., m$. It can be easily shown with minor modification to the proof of Theorem 1 that the weights $\mathbf{w}$ in the bivariate spline representation (2.13) can be approximated with GMRF as well. For example when $\alpha = 2$, the precision matrix of $\mathbf{w}$ is $\mathbf{Q}_2^G(\boldsymbol{\kappa}^2, \boldsymbol{\tau}) = \boldsymbol{\tau}(\boldsymbol{\kappa}^2\mathbf{M}\boldsymbol{\kappa}^2 + 2\boldsymbol{\kappa}^2\mathbf{K} + \mathbf{K}\mathbf{M}^{-1}\mathbf{K})\boldsymbol{\tau}$ or $\mathbf{Q}_2^{LS}(\boldsymbol{\kappa}^2, \boldsymbol{\tau}) = \boldsymbol{\tau}(\boldsymbol{\kappa}^2\mathbf{M}\boldsymbol{\kappa}^2 + 2\boldsymbol{\kappa}^2\mathbf{K} + \mathbf{R})\boldsymbol{\tau}$ for Galerkin or least squares solutions respectively. As stated in Lindgren et al. (2011), by assuming the parameters κ^2 and τ to be constant locally, the solution to the SPDE (2.14) can still be interpreted as a Matérn field over a local area and the associated global non-stationary field can be achieved by combining all the local Matérn fields automatically via the SPDE.

2.3 Numerical simulations

We conduct several numerical simulations to evaluate the performance of the SPDE approach with bivariate splines and compare with the linear finite element approach in Lindgren et al. (2011) in terms of spatial prediction. In all simulations over $\mathbb{R}^2$ we fix $\alpha = 2$ which corresponds to the smoothness parameter $\nu = 1$ in the Matérn covariance function. The Bayesian inference for the model is run with the INLA package (www.r-inla.org) in the statistical computing platform R (R Core Team, 2016). The default priors for the model components in the INLA package are applied. For brevity, our proposed bivariate spline approximation in $S_d^0(\boldsymbol{\Delta})$ is denoted BS-SPDE with $d = 1, 2, ...$ (BS-SPDE-G or BS-SPDE-LS for Galerkin or least squares solution respectively) and the linear finite element approximation is denoted LFE-SPDE.

2.3.1 Comparison of LFE-SPDE and BS-SPDE

Study 1: surfaces with analytical expressions

In this simulation, we compare the LFE-SPDE method and BS-SPDE of degree $d \geq 2$ in data fitting for some common surfaces. Elevations of different surfaces are collected on a grid over square $[-2, 2] \times [-2, 2]$ that is equally spaced every 0.2. Then we make predictions on another finer grid that is equally spaced every 0.01 over square $[-2, 2] \times [-2, 2]$ using the SPDE approach. The prediction accuracy for the whole surface can be measured with mean-square-error $\text{MSE} = \sum_{i=1}^{n}(\hat{f}(\mathbf{u}_i) - f(\mathbf{u}_i))^2/n$, where $f(\mathbf{u}_i)$ is the true elevation on location $\mathbf{u}_i$ and $\hat{f}(\mathbf{u}_i)$ is the prediction using corresponding posterior means.

Four different surfaces are considered: $2\sin(x)\cos(y)$ and $2\exp(-\frac{x^2+y^2}{s})$ with three different shape parameters $s = 2, 1, 0.5$. We construct 35 different meshes that have 2, 3, 6, 13, 19, 28, 53, 96, 112, 148, 212, 279, 342, 390, 444, 520, 705, 874, 1065, 1368, 1802, 2416, 2798, 3176, 3708, 4428, 5514, 6696, 8460, 10958, 15009, 21832, 26718, 33776, 43875 triangles respectively to demonstrate the convergence (mesh size $|\mathbf{\Delta}|$ monotonically decreases roughly from 6.6 to 0.026). For each approach, the associated number of basis functions (denoted by N_B) and CPU time for calling the `inla` programme (denoted by T_{cpu} in seconds) to do the Bayesian inference are recorded when the corresponding MSEs reach levels of 10^{-l}, $l = 1, 2, ..., 8$. N_B is also the dimension of corresponding precision matrix of the weights $\mathbf{w}$ and directly relates to the computational complexity. For example the samples and likelihoods can be computed in $\mathcal{O}(N_B^{3/2})$ operations for two dimensional GMRFs. For comparison, the simulation stops when the number of basis functions of BS-SPDE with $d \geq 2$ exceeds the number of basis functions of LFE-SPDE using the densest mesh. The results are presented in Figure 2.3 where the y-axes for N_B and T_{cpu} are taken on a logarithmic scale.

From Figure 2.3 we can see that in general BS-SPDE with $d \geq 2$ can be more efficient than LFE-SPDE both in terms of number of basis functions and computing time needed to reach specific levels of MSE, especially those lower than 10^{-4}. In the left side of Figure 2.3, the dash lines for BS-SPDE-LS are invisible as they coincide

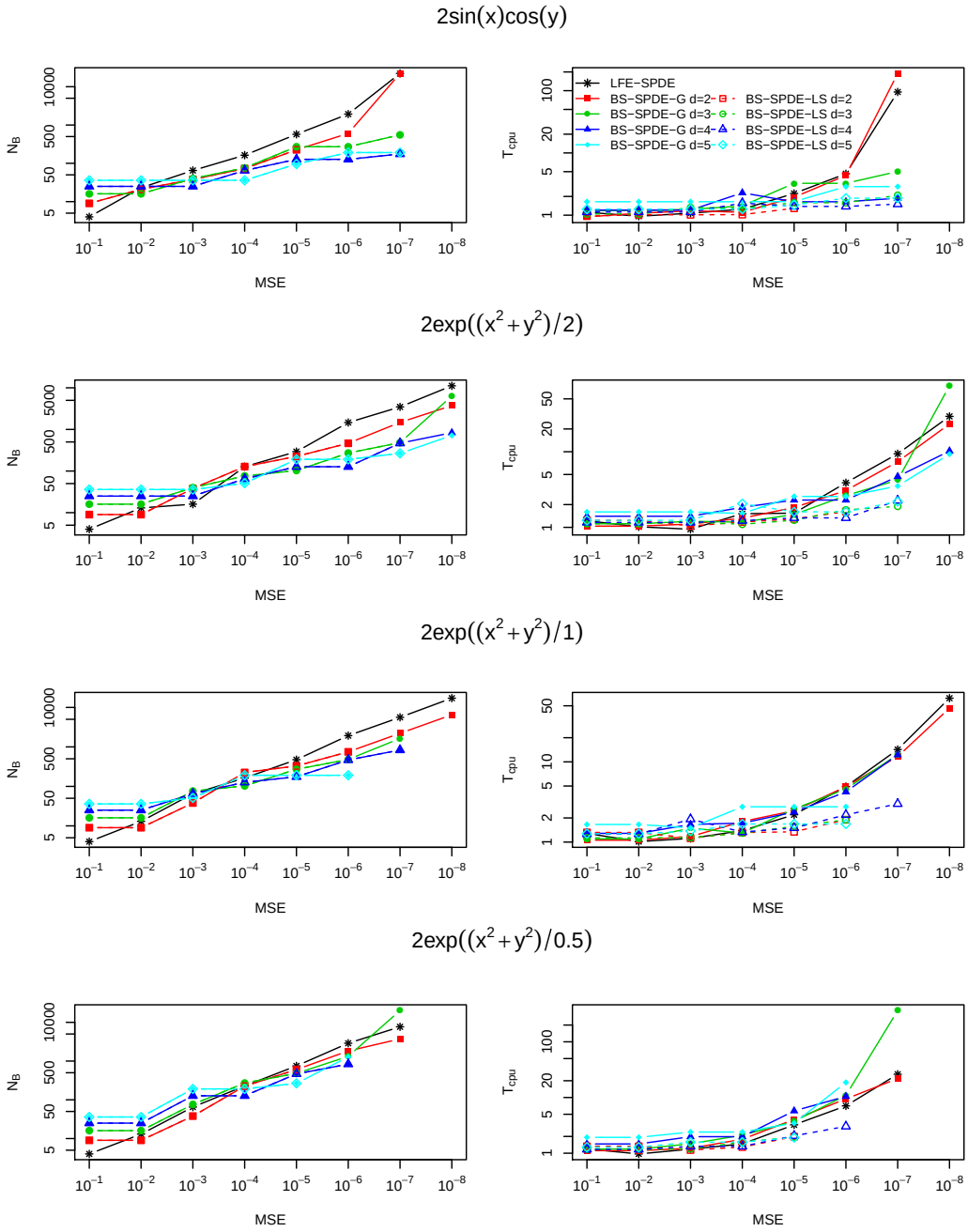


Figure 2.3: Number of basis functions (N_B) and CPU time for `inla` (T_{cpu} in seconds) required by LFE-SPDE, BS-SPDE-G and BS-SPDE-LS with $d = 2, 3, 4, 5$ respectively to reach specific MSE levels for different surfaces.

with the solid lines for BS-SPDE-G, while BS-SPDE-LS is more computationally efficient in general than BS-SPDE-G as shown in the right side. Specifically, for the surface $2\sin(x)\cos(y)$, to reach high precision levels such as 10^{-6} or 10^{-7}, BS-SPDE-G and BS-SPDE-LS with high degree $d \geq 3$ are more efficient since they require only less than 10% of the basis functions and computing time required by LFE-SPDE. For the Gaussian surface $2\exp(-\frac{x^2+y^2}{2})$, BS-SPDE with $d \geq 2$ are generally much more efficient than LFE-SPDE for the MSE levels up to 10^{-7} with about 50% gains in the computing time. But BS-SPDE-LS does not reach the MSE level 10^{-8} and BS-SPDE-G with $d = 3$ takes more computing time than the others to reach the MSE level 10^{-8}. For the next Gaussian shape surface $2\exp(-\frac{x^2+y^2}{1})$ which is steeper than the previous one, BS-SPDE with high degrees can be better than LFE-SPDE for the MSE levels around 10^{-4} to 10^{-6} but their efficiency is decreased to reach the higher precision levels 10^{-7} and 10^{-8}. However, BS-SPDE-LS with $d = 4$ reaches the low MSE level 10^{-7} within only 20% of the computing time required by LEF-SPDE. For the last surface which is quite steep, BS-SPDE-G with $d = 2$ is comparable with LFE-SPDE and reaches the high precision levels 10^{-6}, 10^{-7} by requiring slightly less number of basis functions and similar time, and BS-SPDE-LS with $d = 4$ is more efficient to reach the MSE level 10^{-6}.

From these results, we can conclude that BS-SPDE can be much more efficient in many cases especially when the high precision levels are desired and the target functions are smooth. For functions that are not that smooth, lower degree representations such as LFE-SPDE or BS-SPDE with $d = 2$ might be more appropriate, which is consistent with the general comments by Babuska et al. (1981). Note that even for the last Gaussian shape surface which is much less smooth than the others, BS-SPDE-G with $d = 2$ still can be comparable with LFE-SPDE; and we obtain 50% gains in the computing time using BS-SPDE-LS with $d = 4$ if the MSE level 10^{-6} is desired.

Study 2: bathymetry data

In this study, we compare LFE-SPDE and BS-SPDE in spatial estimation and prediction with real data sets that are extracted from the ETOPO1 Global Relief Model

(Amante and Eakins, 2009), which is a 1 arc-minute global relief model of Earth's surface that integrates land topography and ocean bathymetry. The data is available from National Geophysical Data Center (NGDC), USA. Four different regions around the the Strait of Juan de Fuca area are chosen for this study as shown in Figure 2.4. In general, region 1 covers near shore seabed with relatively simple and gradual variations while the seabed in the other three regions is quite complicated.

Figure 2.4: Four regions extracted from ETOPO1 Global Relief Model around the Strait of Juan de Fuca.

For comparison, both in-sample and out-of-sample predictive fit performance are explored using LFE-SPDE, BS-SPDE-G and BS-SPDE-LS with $d = 2, 3, 4$ based on various meshes. We denote the observations by y_1, y_2, ..., y_n. As for the in-sample fit measurement, root-mean-square-error (RMSE) between the observations and the predictions at the observed locations

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

are calculated where the prediction $\hat{y}_i$ is taken to be the associated posterior mean. Since the SPDE approach aims to estimate the whole surface, smaller RMSE suggests the estimated surface is closer to the measurements at the observed locations. To measure the predictive performance, leave-one-out cross validation is employed

using the embedded function within the INLA package. The logarithmic score (Log Score) of the prediction is defined as

$$\text{Log Score} = -\frac{1}{n}\sum_{i=1}^{n}\log\left[\pi(y_i|y_{-i})\right],$$

where $\pi(y_i|y_{-i})$ is the posterior predictive density of y_i given all the other observations y_{-i}. Therefore, the smaller Log Score is, the more certain we are with the predictions. Six meshes are built as shown in Figure 2.5. The meshes are extended with coarse triangles to avoid boundary effect (Lindgren et al., 2011). The number of basis functions for each combination of mesh and SPDE method is shown in Table 2.1.

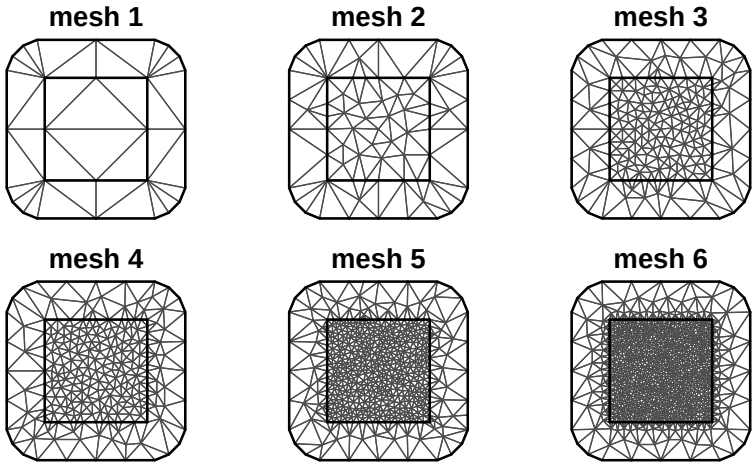


Figure 2.5: Six meshes for Study 2.

Table 2.1: Number of basis functions using LFE-SPDE and BS-SPDE-G/LS with $d = 2, 3, 4$ based on six meshes.

mesh	LFE-SPDE	BS-SPDE $d = 2$	BS-SPDE $d = 3$	BS-SPDE $d = 4$
1	28	89	184	313
2	70	252	547	955
3	195	749	1663	2937
4	244	945	2104	3721
5	536	2111	4726	8381
6	978	3881	8710	15465

Table 2.2 presents the RMSEs and Log Scores using LFE-SPDE, BS-SPDE-G

Table 2.2: Study 2: RMSE and Log Score (RMSE|Log Score) using LFE-SPDE, BS-SPDE-G and BS-SPDE-LS with $d = 2, 3, 4$. The values in bold are obtained based on similar number of basis functions between 945 to 978. The values marked with asterisks $(*)$ are the selected model fit for map reconstruction.

mesh	LFE-SPDE	BS-SPDE-G			BS-SPDE-LS		
		$d=2$	$d=3$	$d=4$	$d=2$	$d=3$	$d=4$
region 1							
1	1.85|2.07	1.41|1.84	1.18|1.73	1.11|1.73	1.41|1.84	1.20|1.72	1.16|1.71
2	1.17|1.71	0.76|1.53	0.55|1.47	**0.29|1.21**	0.88|1.65	0.81|1.70	**0.74|1.73**
3	0.83|2.01	0.46|1.48	0.33|1.42	0.16|1.09	0.43|1.64	0.048|1.18	0.027|0.98*
4	0.79|1.61	**0.46|1.48**	0.34|1.43	0.17|1.16	**0.23|1.43**	0.039|1.08	0.028|0.99
5	0.62|1.51	0.49|1.50	0.42|1.45	0.36|1.38	0.024|1.51	0.022|1.30	0.021|1.08
6	**0.63|1.52**	0.50|1.51	0.49|1.48	0.44|1.53	0.021|1.74	0.020|1.19	0.019|0.92
region 2							
1	79.81|5.82	66.01|5.66	52.26|5.41	47.01|5.30	66.41|5.65	52.21|5.42	45.81|5.27
2	36.20|5.02	21.27|4.53	9.09|3.86	**0.02|0.50**	22.47|4.58	14.52|4.23	**12.48|4.10**
3	17.24|4.34	0.013|0.51	0.0091|0.43	0.0088|0.45	0.015|0.43	0.011|0.41*	0.0094|0.50
4	15.87|4.28	**0.012|0.52**	0.010|0.46	0.010|0.50	**0.013|0.47**	0.0098|0.46	0.0091|0.54
5	0.037|1.85	0.013|0.85	0.010|0.55	0.010|0.48	0.0086|0.61	0.0069|0.55	0.0065|0.57
6	**0.032|1.98**	0.011|0.72	0.010|0.55	0.012|0.53	0.0075|0.71	0.0062|0.64	0.0073|0.81
region 3							
1	41.88|5.18	37.52|5.09	30.83|4.89	25.44|4.69	37.52|5.10	30.51|4.89	25.21|4.68
2	26.64|4.72	11.62|3.93	6.06|3.43	**0.021|0.45**	11.70|3.94	7.83|3.59	**0.029|0.52**
3	11.45|3.95	0.028|0.74	0.016|0.50	0.015|0.48*	0.022|0.58	0.021|0.56	0.017|0.61
4	10.18|3.86	**0.025|0.72**	0.017|0.53	0.015|0.48	**0.021|0.55**	0.019|0.60	0.016|0.59
5	7.21|3.66	0.021|1.01	0.018|0.64	0.016|0.50	0.014|0.69	0.012|0.68	0.012|0.72
6	**0.053|2.08**	0.019|0.99	0.019|0.80	0.019|0.64	0.012|0.82	0.011|0.79	0.012|0.96
region 4							
1	47.12|5.30	35.98|5.07	28.46|4.83	22.95|4.61	36.02|5.07	28.46|4.83	22.73|4.59
2	26.36|4.71	12.56|4.02	8.08|3.71	**0.020|0.48**	12.47|4.01	9.24|3.78	**0.033|0.55**
3	13.85|4.13	0.023|0.62	0.016|0.52	0.013|0.43*	0.019|0.55	0.016|0.54	0.014|0.56
4	12.42|4.04	**0.021|0.76**	0.017|0.54	0.014|0.46	**0.019|0.54**	0.015|0.56	0.014|0.59
5	0.061|2.19	0.019|0.93	0.017|0.64	0.015|0.49	0.012|0.63	0.011|0.63	0.010|0.68
6	**0.049|2.14**	0.017|0.93	0.018|0.78	0.017|0.58	0.011|0.77	0.010|0.77	0.010|0.88

and BS-SPDE-LS with $d = 2, 3, 4$ based on the six meshes respectively. In general, as the triangulation becomes denser, the estimations and predictions are more accurate using both LFE-SPDE and BS-SPDE-G/LS in most cases. For a particular mesh, the RMSEs and Log Scores of BS-SPDE-G/LS with $d \geq 2$ are generally smaller than those of LFE-SPDE. In terms of the number of basis functions, BS-SPDE-G/LS with $d \geq 2$ also demonstrate better performance than LFE-SPDE in most cases. For example, for region 4, BS-SPDE-G with $d = 2$ based on mesh 4 and BS-SPDE-G with $d = 4$ based on mesh 2 yield much smaller RMSEs and Log Scores than LFE-SPDE based on mesh 6 while they have similar numbers of basis functions. In general, BS-SPDE-LS yields smaller RMSEs than BS-SPDE-G in most cases. However, in terms of Log Score, BS-SPDE-G performs better than BS-SPDE-LS in most cases for the other three regions except region 1. We notice the sudden change in the model performance. For example, the RMSE obtained using LFE-SPDE for region 2 is about 15.87 based on mesh 4; it is decreased suddenly to only 0.037 based on mesh 5 or 0.012 using BS-SPDE-G with $d = 2$. This may because the finite elements or splines reach some level of degree of freedom that is enough to model the surface well.

Based on Table 2.2, we can select the models with good performance for continuous map reconstruction among the different combinations of meshes and SPDE approaches. In most cases it is difficult to have a model with the smallest RMSE and Log Score at the same time, so we only choose the one with relatively small RMSE and Log Score. In this way, the reconstructed map can be close to the elevations at the observed locations; meanwhile we are more confident with the predictions at the other locations. As marked with asterisks in Table 2.2, we choose BS-SPDE-LS with $d = 4$ based on mesh 3 for region 1, BS-SPDE-LS with $d = 3$ based on mesh 3 for region 2, BS-SPDE-G with $d = 4$ based on mesh 3 for region 3, and BS-SPDE-G with $d = 4$ based on mesh 3 for region 4. Note that for region 1, BS-SPDE-LS with $d = 4$ based on mesh 6 yields both smallest RMSE and Log Score among all the combinations. However the associated computational cost is much heavier than the others. There is some trade off between model performance and computa-

tional cost. Hence we select the one with relatively good performance and is also computationally efficient. Then the posterior means and standard deviations of the four regions predicted using the respective selected models are displayed in Figure 2.6. The posterior means in general capture the main features of the corresponding regions and the posterior standard deviations provide uncertainty estimates of the predictions. Note that the selection rule of predictive model here is quite simple and subjective. More appropriate model selection techniques can be employed in application.

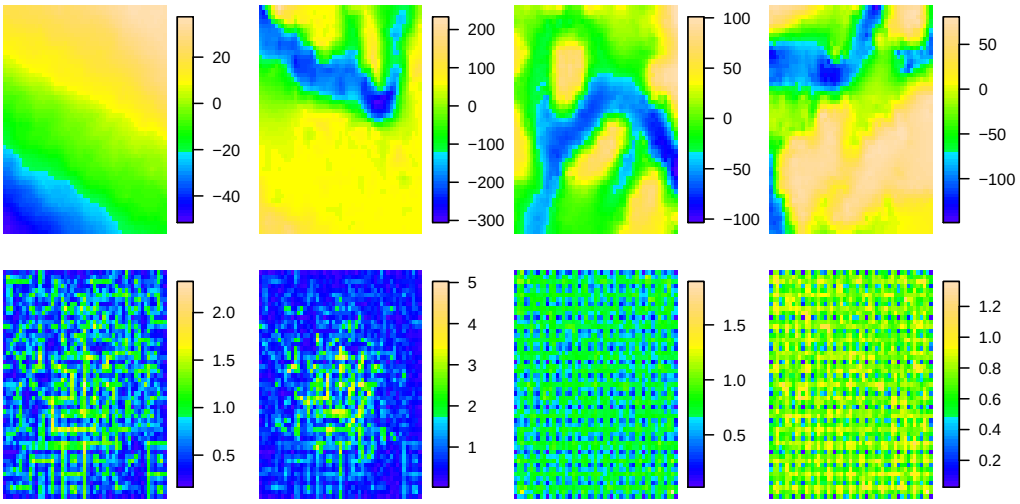


Figure 2.6: Posterior mean (top) and standard deviation (bottom) for the regions 1-4 (left to right), after selection of the appropriate approximation models.

2.3.2 Spatial analysis of ozone levels data over Eastern USA

We analyse a data set of ozone levels at a certain hour in one of days in September, 2005 around the Eastern United States, which is available from the Air Explorer Database of Environmental Protection Agency (EPA), using the non-stationary BS-SPDE-G method. The data set has 546 locations where ozone levels are recorded. As shown in Figure 2.7, the observations of ozone concentration are distributed unevenly and the domain is irregular. Denote the ozone levels by z_i and the associated locations by $\mathbf{s}_i = (x_i, y_i)$ for $i = 1, ..., 546$. We consider a simple spatial model

$$z_i = b_0 + f(\mathbf{s}_i) + \epsilon_i, \quad i = 1, ..., 546,$$

where b_0 is the intercept, $\epsilon_i \sim N(0, \sigma_e^2)$ is i.i.d measurement error and the spatial effect $f(\mathbf{s}_i)$ is assumed to be a non-stationary GF generated by the non-stationary version SPDE (2.14), represented with bivariate splines in $S_d^0(\mathbf{\Delta})$ with $d = 1, 2, 3, 4, 5$. The triangulation $\mathbf{\Delta}$ is shown in Figure 2.7.

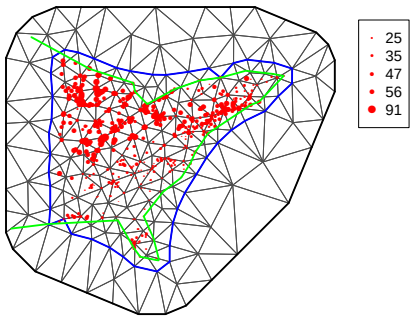


Figure 2.7: Triangulation over Eastern United States; green line: U.S. boundary; red dots: locations of ozone monitoring stations; size proportional to the ozone levels in ppb (parts per billion).

The non-stationary parameters $\tau(\mathbf{u})$ and $\kappa^2(\mathbf{u})$ are represented with two-dimensional B-splines that have n_x and n_y basis functions in the x-direction and y-direction respectively. Therefore at any location $\mathbf{s} = (x, y)$, the basis functions of the associated B-spline can be calculated as $B_{lk}(\mathbf{s}) = B_l^x(x)B_k^y(y)$, where $B_l^x(\cdot)$ and $B_k^y(\cdot)$ are the basis functions in x and y directions respectively, for $l = 1, ..., n_x$ and $k = 1, ..., n_y$. Hence there are $n_x n_y$ basis functions in total for each of the parameters $\kappa^2(\cdot)$ and $\tau(\cdot)$. We consider 12 models A-L with different combinations of the number of basis functions in Table 2.3. Note that with one basis function, the B-spline is constant so model A corresponds to the stationary SPDE model (2.3). The number of basis functions represents the number of basis functions in both x-direction and y-direction; for example in model C, there are 3 basis functions for $\tau(\cdot)$ in each direction which means there are actually $3 \times 3 = 9$ basis functions for $\tau(\cdot)$.

Table 2.3: Number of basis functions for the B-spline in each direction for the parameters.

	A	B	C	D	E	F	G	H	I	J	K	L
$\kappa^2(\cdot)$	1	1	1	1	1	2	3	4	5	2	3	4
$\tau(\cdot)$	1	2	3	4	5	1	1	1	1	2	3	4

To measure the fit and predictive performance and select the appropriate representations for $\kappa^2(\cdot)$ and $\tau(\cdot)$, we employ the leave-one-out cross validation and aim to find the model with the smallest Log Score. Figure 2.8 presents the Log Scores of the 12 models for the two parameters $\kappa^2(\cdot)$ and $\tau(\cdot)$ as shown in Table 2.3 using BS-SPDE-G approach with $d = 1, 2, 3, 4, 5$ respectively.

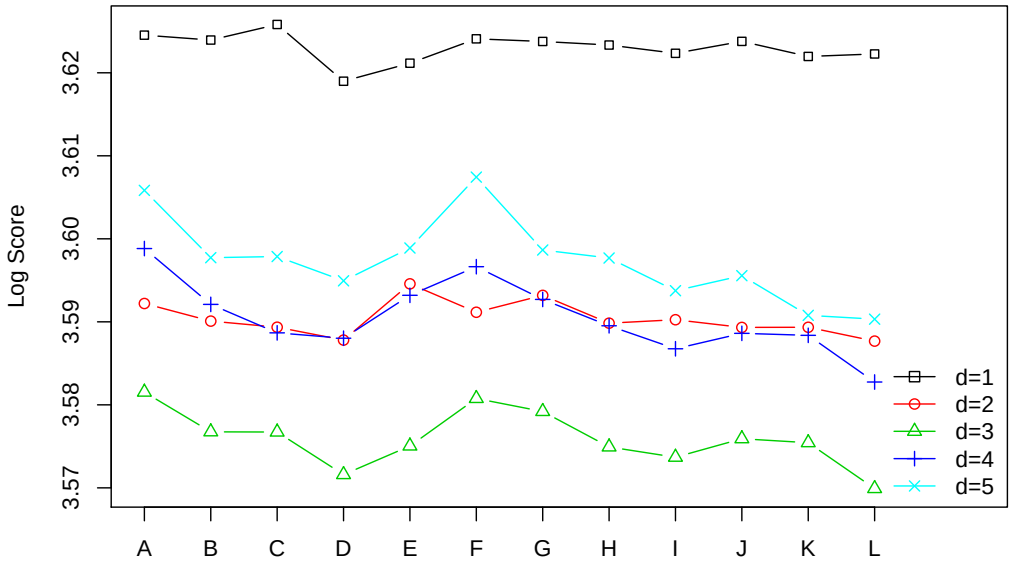


Figure 2.8: Log Scores for the models A-L using BS-SPDE-G with $d = 1, ..., 5$.

It is easy to see that the Log Scores obtained from BS-SPDE-G with higher d are generally smaller than those obtained from BS-SPDE-G with lower d. Using BS-SPDE-G with a specific d, the Log Scores for different representations of $\kappa^2(\cdot)$ and $\tau(\cdot)$ are different. In general the non-stationary models B-L yield smaller Log Scores than the stationary model A. The overall smallest Log Score is obtained with model L and BS-SPDE-G $d = 3$. As shown in Table 2.3, model L corresponds to 4 basis functions for $\kappa^2(\cdot)$ and 4 basis functions for $\tau(\cdot)$ in both x and y direction. This suggests that both $\kappa^2(\cdot)$ and $\tau(\cdot)$ display spatial variation over the domain. The number of parameters may be considered for model selection. We notice that the Log Score obtained using BS-SPDE-G $d = 3$ with model D is only slightly higher than model L while the number of parameters used to represent the non-stationarity is only half of model L. A proper model selection technique would account for it, e.g. AIC and BIC, but this is beyond the scope of this paper.

Then we apply the non-stationary model L and predict ozone levels using the BS-SPDE-G approach with $d = 3$. Figure 2.9 displays the posterior mean and standard deviation of the predictions given the observations presented in Figure 2.7. As we can see, the predicted ozone level is low in the south-east corner and at the top of the north-east corner and high in the north and middle area, which is consistent with the observations. Furthermore, the posterior predictive standard deviation shows some spatial variation over the entire domain because of the irregular distribution of the observations and the non-stationarity of the SPDE model.

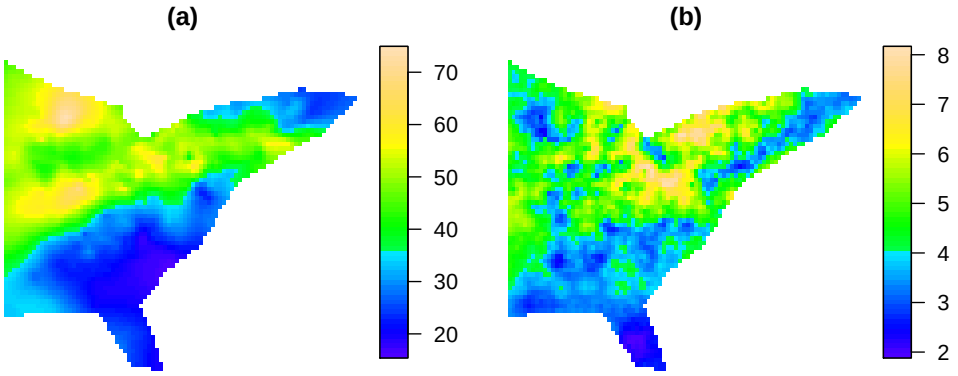


Figure 2.9: (a) Posterior mean; (b) posterior standard deviation: ozone levels over Eastern United States predicted using BS-SPDE-G with $d = 3$ and non-stationary model L.

2.4 Discussion

We have shown that higher order polynomial basis can be easily implemented in the SPDE framework for GFs using bivariate splines. Both the theoretical results and numerical simulations have demonstrated the advantages of this new approach over the linear finite element approach in terms of approximation accuracy and computational efficiency. By using higher degree representations, we can also implement the least squares solutions to the SPDE (2.3) for $\alpha = 2$. This is more computationally efficient than the corresponding Galerkin solutions due to the sparser structures. We have shown that the SPDE approach can be applied to the spatial modelling of bathymetry. The current commonly used mapping tools, e.g. Generic Mapping Tools (GMT) used by NOAA (Eakins and Taylor, 2010), do not include uncertainty

estimates of the maps. GMT also requires high smoothness conditions, see Smith and Wessel (1990) and Wessel and Bercovici (1998), that may not be appropriate for the bathymetry and topography. Hence, the computationally efficient SPDE approach has promising potential in spatial mapping.

There is still room for further investigation in the SPDE approach with bivariate splines. It has been suggested by the numerical simulations that the degree of polynomial basis has an impact on the performance of the SPDE approach. Thus it is essential to choose appropriate degrees. In fact, the degree of bivariate splines can be adaptive. Hu et al. (2007) proposed a new spline method which allows automatic degree raising over triangles of interest. This new method is able to solve linear PDEs very effectively and efficiently. Another extension to manifolds could be considered. Lai et al. (2009) discussed the application of spherical splines in geopotential approximation where the techniques of triangulated spherical splines can be applied to represent the Matérn fields on manifolds. Furthermore, when α is larger than 2, which means the smoothness parameter ν in (2.2) increases as well, sample paths of the Matérn fields are smoother (Paciorek and Schervish, 2004). In this case, smoother representations of the GFs are desired. However, it is quite difficult to implement higher orders of smoothness in conventional finite elements. But within the bivariate splines framework, higher orders of smoothness conditions can be implemented easily by imposing linear constraints on the B-coefficients (Lai and Schumaker, 2007). However, the implementation within the SPDE framework is non-trivial and needs to be investigated as the large number of linear constraints are computationally expensive.

Chapter 3

Dimension Reduction for Emulation: Application to the Influence of Bathymetry on Tsunami Heights

3.1 Introduction

Simulators are widely employed to reproduce physical processes and explore their behaviour, in fields such as fluid dynamics or climate modelling. To characterise the impact of the uncertainties in the boundary conditions or the parameterisations of the underlying physical processes, a sufficient number of simulations are required. However, when the simulators are computationally expensive, as it is the case for high accuracy simulations, the task can become extremely costly or even prohibitive. One prevailing way to overcome this hurdle is to construct statistical surrogates, namely emulators, to approximate the computer simulators in a probabilistic way (Sacks et al., 1989). Emulators are trained on a relatively small number of well-chosen simulations, i.e. a design of computer experiments. Outputs at any input can be predicted at little computational cost with emulators. One can then employ emulators for any subsequent purposes such as uncertainty propagation, sensitivity analysis and calibration.

With high-dimensional inputs, say beyond 20 dimensions, one would need to use a large design to explore the input space, typically in the order of 10 times the

number of dimensions, for a reasonable level of approximation. One would face serious computational problem since the original simulator cannot be run many times. Advanced designs such as Latin Hypercubes or new sequential designs (Beck and Guillas, 2016) that are more efficient than Latin Hypercubes only alleviate the issue. As a result, methods that adequately reduce the dimension of the input space are required, as high-dimensional inputs are often present in computer models, e.g. as boundary conditions like the bathymetry in tsunami modelling. Some approaches ignore high-dimensional inputs and add stochastic terms to account for their contribution (Iooss and Ribatet, 2009, Marrel et al., 2012). These methods are easy to implement and effective in some applications. However, repeated simulations at the same input parameters that are encoded in the emulation are often required to estimate the variability due to those parameters that are ignored. The variability estimates are often restricted to the second moments, and the input-output relationships over the ignored inputs are not clear. Constantine et al. (2014) proposed to find rotations of the input space with the strongest variability in the gradients of the simulators and constructed a response surface on such a low-dimensional active subspace. This Active Subspace (AS) method has been demonstrated to be effective theoretically and numerically. Constantine and Gleich (2015) studied further the properties of the Monte Carlo approximation of the subspace. However, this method requires the calculation of a sufficient number of gradients explicitly, which unfortunately prevents its use in many applications. The gradients are often unavailable in many realistic simulators, and typically intractable for systems of mixed PDEs or multi-physics simulations. Even in the rare situations where gradients are computable numerically, the computational cost of obtaining them could be prohibitive.

The concept of active subspace is closely related to the sufficient dimension reduction (SDR) (Cook, 1994, 2009) and effective dimension reduction (EDR) (Li, 1991) in the statistical community. Given an explanatory variable $\mathbf{X} \in \mathbb{R}^m$ (input) and response variable Y (output), the aim of SDR (or EDR) is to find the directions in the subspace of $\mathbf{X}$ that contain sufficient information about the response for

statistical inference. More specifically, a SDR $R(\mathbf{X}) \in \mathbb{R}^d$ where $d < m$ satisfies $p(Y|\mathbf{X}) = \tilde{p}(Y|R(\mathbf{X}))$, where $p(Y|\mathbf{X})$ and $\tilde{p}(Y|R(\mathbf{X}))$ are conditional probability density functions with respect to $\mathbf{X}$ and $R(\mathbf{X})$ respectively. The EDR approach aims to specifically find a linear projection matrix $\mathbf{B}$ onto a d-dimensional subspace $(d < m)$ such that $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$ and

$$p(Y|\mathbf{X}) = \tilde{p}(Y|\mathbf{B}^T\mathbf{X}) \quad \text{or equivalently} \quad Y \perp X|\mathbf{B}^T X. \tag{3.1}$$

Several methods have been developed to find SDR including nonparametric approaches such as sliced inverse regression (SIR) (Li, 1991), minimum average variance estimation (MAVE) (Xia et al., 2002), and parametric approaches like principal fitted components (PFC) (Cook, 2007, Cook and Forzani, 2008). In this paper, we adopt the gradient-based kernel dimension reduction (gKDR) developed by Fukumizu and Leng (2014) to construct low-dimensional approximations to the simulators. The gKDR method does not require any strong assumptions on the variables and distributions. The response variable can be of arbitrary type: continuous or discrete, univariate or multivariate. Unlike the active subspace method in Constantine et al. (2014), gradients are not required to be computed explicitly but are estimated non-parametrically and implicitly using stable kernel methods. The gKDR approach ends up with an eigen-problem without any needs of elaborate numerical optimisation and thus can be applied to large and high-dimensional problems. Our proposed approach therefore provides an answer to the dimension reduction issue in emulation for a wide range of problems that cannot be tackled using existing methods at the moment.

We introduce a joint framework to approximate the high-dimensional simulators by building statistical emulators within the gKDR approach. Deterministic simulators are considered here, however, the framework could potentially be applied to stochastic simulators, with additional treatments to the stochastic effect in the emulation, see e.g. Henderson et al. (2009). Throughout this chapter, the mainstream Gaussian process (GP) emulators are employed for illustration. But the general framework and most of the results in this paper would hold potentially for

other emulation techniques.

The chapter is organised as follows. Section 3.2 and 3.3 review GP emulator and the gKDR approach respectively. In Section 3.4, a joint framework of dimension reduction combined with emulation is proposed and some theoretical properties are established. Section 3.6 contains the numerical applications to an elliptic PDE and to the propagation of uncertainties in the bathymetry to tsunami wave heights as well as comparison with other existing methods.

3.2 Gaussian process emulator

A Gaussian process is a collection of random variables such that any finite subset of these variables follow a joint Gaussian distribution (Rasmussen and Williams, 2006). It is widely used in various scientific fields. Here we briefly review some basics of its application in statistical emulation.

A deterministic simulator with multivariate input $\mathbf{X} = (x_1, ..., x_m)^T \in \mathbb{R}^m$ and univariate output $Y \in \mathbb{R}$ can be represented as $Y = f(\mathbf{X})$. The GP emulator assumes that the simulator output $Y = f(\mathbf{X})$ can be modelled with a Gaussian process. It is commonly assumed that the mean $\mathrm{E}(Y) = m(\mathbf{X}) = \mathbf{h}^T(\mathbf{X})\boldsymbol{\beta}$, where $\mathbf{h}(\mathbf{X})$ is a q-vector of pre-defined regression functions and the coefficients $\boldsymbol{\beta} \in \mathbb{R}^q$. In practice, a constant or linear form for the regression functions would perform well. The covariance between two simulator outputs $Y = f(\mathbf{X})$ and $Y' = f(\mathbf{X}')$ is usually represented as $\mathrm{Cov}(Y, Y') = k(\mathbf{X}, \mathbf{X}') = \sigma^2 c(\mathbf{X}, \mathbf{X}')$, where the positive scalar parameter σ^2 is the process variance and $c(\mathbf{X}, \mathbf{X}')$ is the correlation function. A common choice for the correlation function is the squared-exponential correlation $c(\mathbf{X}, \mathbf{X}') = \prod_{i=1}^{m} \exp\left(-(x_i - x_i')^2/\delta_i^2\right)$, where $\boldsymbol{\delta} = (\delta_1, ..., \delta_m)^T \in (0, \infty)^m$ controls the correlation lengths.

Suppose the simulator is run at n inputs $\mathbf{X}_1, ..., \mathbf{X}_n$ and the respective outputs are $Y_1, ..., Y_n$. Firstly, let us consider the case that $m(\mathbf{X}) = 0$ and $k(\mathbf{X}, \mathbf{X}') = k(\mathbf{X}, \mathbf{X}'; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ contains the set of hyperparameters relating to the covariance function. At any n^* desired inputs $\mathbf{X}_1^*, \mathbf{X}_2^*, ..., \mathbf{X}_{n^*}^*$, the respective outputs are denoted by $Y_1^*, Y_2^*, ..., Y_{n^*}^*$. Then the joint distribution of $\mathbf{Y} = (Y_1, ..., Y_n)^T$ and

$\mathbf{Y}^* = (Y_1^*, ..., Y_{n^*}^*)$, conditional on the covariance function $k(\cdot, \cdot)$, are Gaussian as follows,

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}^* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{K}_*^T \\ \mathbf{K}_* & \mathbf{K}_{**} \end{bmatrix}\right),$$

where $\mathbf{K}$, $\mathbf{K}_*$ and $\mathbf{K}_{**}$ are $n \times n$, $n^* \times n$ and $n^* \times n^*$ matrices respectively with the associated (i, j)-th entry as $\mathbf{K}(i, j) = k(\mathbf{X}_i, \mathbf{X}_j)$, $\mathbf{K}_*(i, j) = k(\mathbf{X}_i^*, \mathbf{X}_j)$ and $\mathbf{K}_{**}(i, j) = k(\mathbf{X}_i^*, \mathbf{X}_j^*)$. Conditioning the Gaussian distribution on the observed data and covariance function, we have

$$\mathbf{Y}^* | \mathbf{Y}, k(\cdot, \cdot; \boldsymbol{\theta}) \sim \mathrm{N}(\mathbf{K}_* \mathbf{K}^{-1} \mathbf{Y}, \mathbf{K}_{**} - \mathbf{K}_* \mathbf{K}^{-1} \mathbf{K}_*^T).$$

The output at any desired input predicted using GP emulator is a distribution rather than a single value. This could be used to estimate the uncertainty introduced into the prediction with emulator and to evaluate the confidence about the prediction. To complete the prediction, we have to specify the hyperparameters $\boldsymbol{\theta}$ from the covariance function $k(\cdot, \cdot; \boldsymbol{\theta})$ properly. It is possible to make a fully Bayesian inference with appropriate prior $\pi(\boldsymbol{\theta})$. But this usually requires costly MCMC approach for the analytically intractable posterior. In practice, a computationally cheap alternative is often employed by specifying the hyperparameters $\boldsymbol{\theta}$ at the most probable values. This could be done by maximising the marginal likelihood. Observing that $\mathbf{Y} \sim \mathrm{N}(\mathbf{0}, \mathbf{K}(\boldsymbol{\theta}))$, we have the log marginal likelihood,

$$L(\boldsymbol{\theta}) = \log p(\mathbf{Y}|\boldsymbol{\theta}) = -\frac{1}{2} \mathbf{Y}^T \mathbf{K}^{-1} \mathbf{Y} - \frac{1}{2} \log |\mathbf{K}| - \frac{n}{2} \log 2\pi.$$

Then the hyperparameters can be estimated by $\hat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$.

When we assume the non-zero mean trend $m(\mathbf{X}) = \mathbf{h}^T(\mathbf{X})\boldsymbol{\beta}$, a prior for the parameter $\boldsymbol{\beta}$ may be imposed. One of the popular choices is a Gaussian prior, $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{b}, \mathbf{V})$, which forms a conjugate prior with the GP likelihood. Then we have the following GP model,

$$f(\mathbf{X}) \sim GP(\mathbf{h}^T(\mathbf{X})\mathbf{b}, k(\mathbf{X}, \mathbf{X}') + \mathbf{h}^T(\mathbf{X})\mathbf{V}\mathbf{h}(\mathbf{X}')).$$

The prediction process can be obtained following the similar fashion as in the zero mean case:

$$\mathbf{Y}^*|\mathbf{Y}, k(\cdot,\cdot;\boldsymbol{\theta}) \sim \mathrm{N}(\hat{\mathbf{m}}^*, \hat{\boldsymbol{\Sigma}}^*),$$

with

$$\hat{\mathbf{m}}^* = \mathbf{H}^{*T}\hat{\boldsymbol{\beta}} + \mathbf{K}_*\mathbf{K}^{-1}(\mathbf{Y} - \mathbf{H}^T\hat{\boldsymbol{\beta}}),$$

and

$$\hat{\boldsymbol{\Sigma}}^* = \mathbf{K}_{**} + \mathbf{P}^T(\mathbf{V}^{-1} + \mathbf{H}\mathbf{K}^{-1}\mathbf{H}^T)^{-1}\mathbf{P},$$

where $\mathbf{H} = (\mathbf{h}(\mathbf{X}_1), ..., \mathbf{h}(\mathbf{X}_n))$, $\mathbf{H}^* = (\mathbf{h}(\mathbf{X}_1^*), ..., \mathbf{h}(\mathbf{X}_{n^*}^*))$,$\mathbf{P} = \mathbf{H}^* - \mathbf{H}\mathbf{K}^{-1}\mathbf{K}_*^T$, and $\hat{\boldsymbol{\beta}} = (\mathbf{V}^{-1} + \mathbf{H}\mathbf{K}^{-1}\mathbf{H}^T)^{-1}(\mathbf{H}\mathbf{K}^{-1}\mathbf{Y} + \mathbf{V}^{-1}\mathbf{b})$. The hyperparameters in the covariance function can also be estimated by maximising the marginal likelihood,

$$\begin{aligned}\log p(\mathbf{Y}|\mathbf{b}, \mathbf{V}) = &-\frac{1}{2}(\mathbf{H}^T\mathbf{b} - \mathbf{Y})^T(\mathbf{K} + \mathbf{H}^T\mathbf{V}\mathbf{H})^{-1}(\mathbf{H}^T\mathbf{b} - \mathbf{Y}) \\ &- \frac{1}{2}\log|\mathbf{K} + \mathbf{H}^T\mathbf{V}\mathbf{H}| - \frac{n}{2}\log 2\pi.\end{aligned}$$

Usually there is no sufficient information about the parameter $\boldsymbol{\beta}$, hence a vague prior can be imposed by letting $\mathbf{V}^{-1} \to \mathbf{O}$ and $\mathbf{b} = \mathbf{0}$, where $\mathbf{O}$ is the matrix of zeros. In this case, the conditional predictive process can be updated as

$$\mathbf{Y}^*|\mathbf{Y}, k(\cdot,\cdot;\boldsymbol{\theta}) \sim \mathrm{N}(\hat{\mathbf{m}}^*, \hat{\boldsymbol{\Sigma}}^*),$$

with

$$\hat{\mathbf{m}}^* = \mathbf{H}^{*T}\hat{\boldsymbol{\beta}} + \mathbf{K}_*\mathbf{K}^{-1}(\mathbf{Y} - \mathbf{H}^T\hat{\boldsymbol{\beta}}),$$

and

$$\hat{\boldsymbol{\Sigma}}^* = \mathbf{K}_{**} + \mathbf{P}^T(\mathbf{H}\mathbf{K}^{-1}\mathbf{H}^T)^{-1}\mathbf{P},$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{H}\mathbf{K}^{-1}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{K}^{-1}\mathbf{Y}$. This is closely related to the t-process as described in O'Hagan (1994) when a weak prior for $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta})$ that $\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}) \propto \sigma^{-2}\pi_{\boldsymbol{\delta}}(\boldsymbol{\delta})$ is assumed with the mean function $m(\cdot) = \mathbf{h}^T(\cdot)\boldsymbol{\beta}$ and the covariance function $k(\cdot,\cdot) = \sigma^2 c(\cdot,\cdot;\boldsymbol{\delta})$, where $\boldsymbol{\delta}$ contains the parameters in the correlation

function $c(\cdot, \cdot)$.

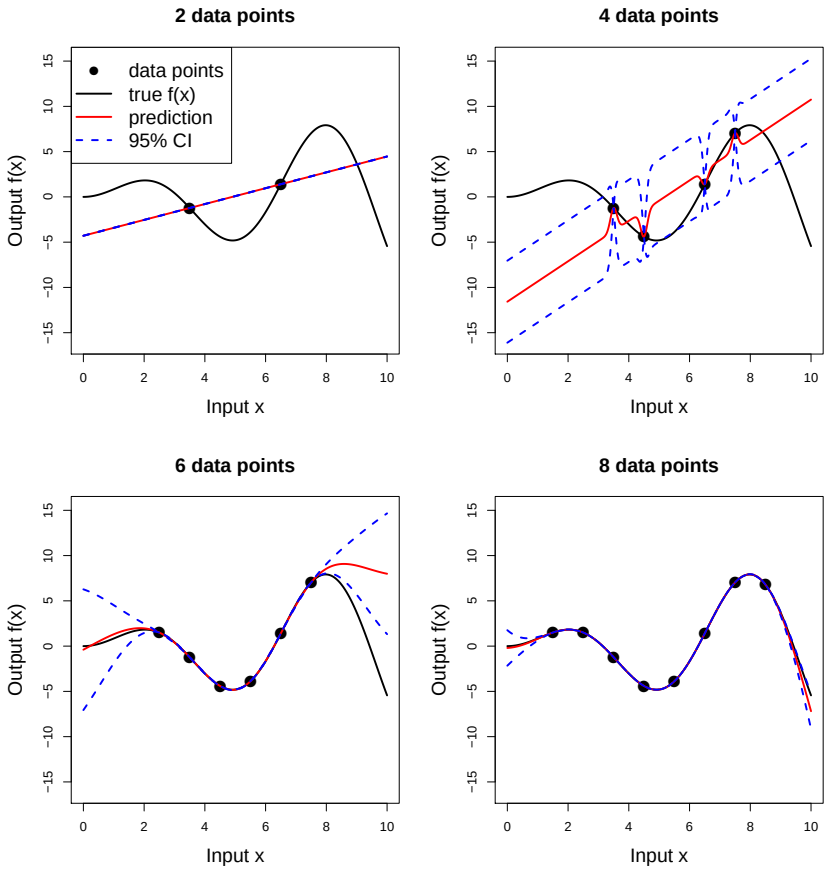


Figure 3.1: An example of Gaussian process emulator for a simple function $f(x) = x\sin(x)$ with increasing number of training data.

When $k(\mathbf{X}, \mathbf{X}') = \sigma^2 c(\mathbf{X}, \mathbf{X}')$ is used with a continuous correlation function $c(\cdot, \cdot)$, such as the squared-exponential correlation, the emulator interpolates through the training data points, i.e. $\hat{m}(\mathbf{X}_i) = Y_i$ and $\hat{v}(\mathbf{X}_i) = 0$ at the training points $\{\mathbf{X}_i\}_{i=1}^n$. Figure 3.1 demonstrates an application of GP emulator in the prediction of a univariate function $f(x) = x\sin(x)$ with different number of training data points. The prediction goes through exactly the training data points with a zero width of the 95% confidence interval. When a nugget term is included, this is no longer true. A nugget term can be included, e.g. to mitigate numerical instabilities or account for the stochastic terms in simulations (Andrianakis and Challenor, 2012). The correlation function $c(\mathbf{X}, \mathbf{X}')$ can be extended with the addition of a nugget as $\tilde{c}(\mathbf{X}, \mathbf{X}') = \nu I_{\mathbf{X}=\mathbf{X}'} + (1-\nu)c(\mathbf{X}, \mathbf{X}')$, where $\nu > 0$ is the nugget term,

and $I_{\mathbf{X}=\mathbf{X}'}$ is the indicator function that is 1 if $\mathbf{X} = \mathbf{X}'$ and 0 otherwise. The associated correlation matrix is $\tilde{\mathbf{K}} = (1-\nu)\mathbf{K} + \nu\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

In practice, the error in the prediction of a GP emulator depends on the number of training data points. As there are more and more training data points, the GP emulator is expected to recover the simulator. This trend is also illustrated in Figure 3.1 where the prediction is closer to the true function when the emulator is trained with more data points. There are various theoretical results on how well the GP emulator $\hat{f}$ can approximate the simulator f in the literature. For example, given n training samples that are quasi-uniformly distributed on $\Omega \subset \mathbb{R}^d$, the error can be bounded (Fasshauer, 2011) as $\|f - \hat{f}\|_\infty \leq C_d n^{-p/d} \|f\|_{\mathcal{H}}$ for any f in some proper space $\mathcal{H}$. This result suggests that $\hat{f}$ provides arbitrarily high approximation order when $p = \infty$, i.e. f is infinitely smooth. However, this rate decreases as the dimension increases and the constant C_d also grows with d. This implies that more evaluations of the simulator are required to train an accurate emulator when the number of input parameters d increases and the associated computational cost could increase dramatically. Therefore, it is desirable to reduce the dimension of the problem from the perspectives of both accuracy and efficiency.

3.3 Gradient-based kernel dimension reduction

For a set Ω, a symmetric kernel $k : \Omega \times \Omega \to \mathbb{R}$ is positive-definite if $\sum_{i,j=1}^{n} c_i c_j k(\omega_i, \omega_j) \geq 0$ for any $\omega_1, ..., \omega_n \in \Omega$ and $c_1, ..., c_n \in \mathbb{R}$. Then a positive-definite kernel k on Ω is uniquely associated with a Hilbert space $\mathcal{H}$ consisting of functions on Ω such that (i) $k(\cdot, \omega) \in \mathcal{H}$; (ii) the linear hull of $\{k(\cdot, \omega) | \omega \in \Omega\}$ is dense in $\mathcal{H}$; (iii) $\langle h, k(\cdot, \omega)\rangle_{\mathcal{H}} = h(\omega)$ for any $\omega \in \Omega$ and $h \in \mathcal{H}$ where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$. Because the third property implies that the kernel k reproduces any function $h \in \mathcal{H}$, the Hilbert space $\mathcal{H}$ is called the reproducing kernel Hilbert space (RKHS) associated with k. Let $(\mathbf{X}, Y)$ be a random vector on the domain $\mathbb{R}^m \times \mathcal{Y}$, and $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be positive definite-kernels on $\mathbb{R}^m$ and $\mathcal{Y}$ with respective RKHS $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$. We shortly present the salient facts about the gKDR method.

If there is some $m \times d$ matrix $\mathbf{B}$ with $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$ satisfying the EDR condition

(3.1), Fukumizu and Leng (2014) note that for any $g \in \mathcal{H}_\mathcal{Y}$, there exists a function $\varphi_g(\mathbf{z})$ on $\mathbb{R}^d$ such that

$$\mathrm{E}\left[g(Y)|\mathbf{X}\right] = \varphi_g(\mathbf{B}^T\mathbf{X}).$$

Then for any $\mathbf{X} = \mathbf{x}$, under mild assumptions, we have,

$$\frac{\partial}{\partial x_i}\mathrm{E}\left[g(Y)|\mathbf{X} = \mathbf{x}\right] = \sum_{a=1}^{d}\mathbf{B}_{ia}\langle g, \nabla_a\varphi(\mathbf{B}^T\mathbf{x})\rangle_{\mathcal{H}_\mathcal{Y}}.$$

On the other hand, defining the cross-covariance operator $C_{Y\mathbf{X}} : \mathcal{H}_\mathcal{X} \to \mathcal{H}_\mathcal{Y}$ as the operator such that

$$\langle h_2, C_{Y\mathbf{X}}h_1\rangle_{\mathcal{H}_\mathcal{Y}} = \mathrm{E}\left[h_1(\mathbf{X})h_2(Y)\right]$$

holds for all $h_1 \in \mathcal{H}_\mathcal{X}$, $h_2 \in \mathcal{H}_\mathcal{Y}$, and using the fact that

$$C_{\mathbf{X}\mathbf{X}}\,\mathrm{E}[g(Y)|\mathbf{X}] = C_{\mathbf{X}Y}g$$

if $\mathrm{E}[g(Y)|\mathbf{X}] \in \mathcal{H}_\mathcal{X}$ for any $g \in \mathcal{H}_\mathcal{Y}$ (Fukumizu et al., 2004), we obtain

$$\frac{\partial}{\partial x_i}\mathrm{E}\left[g(Y)|\mathbf{X} = \mathbf{x}\right] = \left\langle g, C_{Y\mathbf{X}}C_{\mathbf{X}\mathbf{X}}^{-1}\frac{\partial k_{\mathcal{X}(\cdot,\mathbf{x})}}{\partial x_i}\right\rangle_{\mathcal{H}_\mathcal{Y}}.$$

Equating the two expressions above yields for $i, j = 1, ..., m$,

$$\begin{aligned}\mathbf{M}_{ij}(\mathbf{x}) &= \left\langle C_{Y\mathbf{X}}C_{\mathbf{X}\mathbf{X}}^{-1}\frac{\partial k_{\mathcal{X}(\cdot,\mathbf{x})}}{\partial x_i}, C_{Y\mathbf{X}}C_{\mathbf{X}\mathbf{X}}^{-1}\frac{\partial k_{\mathcal{X}(\cdot,\mathbf{x})}}{\partial x_j}\right\rangle_{\mathcal{H}_\mathcal{Y}} \\ &= \sum_{a,b=1}^{d}\mathbf{B}_{ia}\mathbf{B}_{jb}\langle\nabla_a\varphi(\mathbf{B}^T\mathbf{x}), \nabla_b\varphi(\mathbf{B}^T\mathbf{x})\rangle_{\mathcal{H}_\mathcal{Y}}.\end{aligned}$$

Therefore, the dimension reduction projection matrix $\mathbf{B}$ is formed as the eigenvectors associated with the nontrivial eigenvalues of the $m \times m$ matrix $\mathbf{M}(\mathbf{x})$.

Given i.i.d. samples $(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)$, Fukumizu and Leng (2014) proposed to estimate the $\mathbf{B}$ matrix with $\tilde{\mathbf{B}}$ that contains the first d eigenvectors of the

following $m \times m$ symmetric matrix,

$$\tilde{\mathbf{M}}_n = \frac{1}{n}\sum_{i=1}^{n} \nabla \mathbf{k}_{\mathbf{X}}(\mathbf{X}_i)^T (\mathbf{G}_{\mathbf{X}} + n\epsilon_n \mathbf{I})^{-1} \mathbf{G}_Y (\mathbf{G}_{\mathbf{X}} + n\epsilon_n \mathbf{I})^{-1} \nabla \mathbf{k}_{\mathbf{X}}(\mathbf{X}_i),$$

where $\mathbf{G}_{\mathbf{X}}$ and $\mathbf{G}_Y$ are the Gram matrices with the (i,j)-entry as $k_{\mathcal{X}}(\mathbf{X}_i, \mathbf{X}_j)$ and $k_{\mathcal{Y}}(Y_i, Y_j)$ respectively, $\nabla \mathbf{k}_{\mathbf{X}}(\cdot) = (\partial k_{\mathcal{X}}(\mathbf{X}_1, \cdot)/\partial \mathbf{x}, ..., \partial k_{\mathcal{X}}(\mathbf{X}_n, \cdot)/\partial \mathbf{x})^T \in \mathbb{R}^{n \times m}$.

Sometimes there may not exist such a sufficient subspace rigorously so that $d = m$, or we may want to select less dimensions $d' < d$ for later analysis even in cases where such a subspace exists in order to achieve a more stringent reduction (albeit with a small loss). For convenience, we slightly reformulate the gKDR approach into a more general form which does not change the results in Fukumizu and Leng (2014). Let $\mathbf{W}$ be an $m \times m$ matrix with $\mathbf{W}^T\mathbf{W} = \mathbf{I}_m$, satisfying $p(Y|\mathbf{X}) = \tilde{p}(Y|\mathbf{W}^T\mathbf{X})$. In fact, if there exists a $\mathbf{B}$ matrix satisfying (3.1), we can just set $\mathbf{W} = [\mathbf{B}\ \mathbf{C}]$, where $\mathbf{C}$ is an $m \times (m-d)$ matrix such that $\mathbf{C}^T\mathbf{C} = \mathbf{I}_{m-d}$ and the column vectors of $\mathbf{C}$ are orthogonal to those of $\mathbf{B}$; otherwise, $\mathbf{W} = \mathbf{B}$ and $d = m$.

Following the same procedure as before, it is easy to see that

$$\mathbf{M}_{ij}(\mathbf{x}) = \sum_{a,b=1}^{m} \mathbf{W}_{ia}\mathbf{W}_{jb} \langle \nabla_a \varphi(\mathbf{W}^T\mathbf{x}), \nabla_b \varphi(\mathbf{W}^T\mathbf{x}) \rangle_{\mathcal{H}_{\mathcal{Y}}}.$$

If there exists $\mathbf{B}$ satisfying (3.1) with $d < m$, $\nabla_a \varphi(\mathbf{W}^T\mathbf{x}) = 0$ for any $a > d$, hence the respective columns correspond to the zero eigenvalues of $\mathbf{M}(\mathbf{x})$. The projection matrix $\mathbf{W}$ does not depend on the value of $\mathbf{x}$, while the nontrivial eigenvalues vary with $\mathbf{x}$. Therefore, we obtain the following eigen-decomposition

$$\mathbf{M}(\mathbf{x}) = \mathbf{W}\mathbf{\Lambda}(\mathbf{x})\mathbf{W}^T,\ \mathbf{\Lambda}(\mathbf{W}) = \mathrm{diag}(\lambda_1(\mathbf{x}), ..., \lambda_m(\mathbf{x})). \tag{3.2}$$

3.4 Joint emulation with dimension reduction

The gKDR approach is now applied, together with GP emulation, to construct a low-dimensional approximation to a simulator. Thus the following procedure is employed to emulate a high-dimensional simulator.

Step 1. Given a set of n_1 simulator's runs $(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_{n_1}, Y_{n_1})$, estimate the projection matrix $\tilde{\mathbf{W}}$ using the gKDR approach.

Step 2. Split $\tilde{\mathbf{W}}$ into $[\tilde{\mathbf{W}}_1\ \tilde{\mathbf{W}}_2]$, where $\tilde{\mathbf{W}}_1$ consists of the first d columns of $\tilde{\mathbf{W}}$ corresponding to the largest d eigenvectors.

Step 3. Design a set of n_2 runs $(\mathbf{X}'_1, Y'_1), ..., (\mathbf{X}'_{n_2}, Y'_{n_2})$ of the simulator, e.g. based on the reduced space $\tilde{\mathbf{W}}_1^T\mathbf{X}$, and construct an emulator using the lower dimensional pairs $(\tilde{\mathbf{W}}_1^T\mathbf{X}'_1, Y'_1), ..., (\tilde{\mathbf{W}}_1^T\mathbf{X}'_{n_2}, Y'_{n_2})$.

In Step 1, sufficient samples are needed to estimate $\tilde{\mathbf{W}}$ accurately. The theoretical results in Fukumizu and Leng (2014) on the convergence rate of $\tilde{\mathbf{M}}_n$ would provide some insights. In practice, the number of directions that have a major influence may also affect the sample size n_1 needed. Step 2 requires an appropriate selection of d to construct an efficient and effective emulator. The samples to train the emulator in Step 3 can be different (e.g. additional runs) from those already collected to find $\tilde{\mathbf{W}}$ in Step 1. There is a benefit in terms of design arising from the dimension reduction. Indeed, in step 3, the design can be built to explore the reduced space of possible $\tilde{\mathbf{W}}_1^T\mathbf{X}'$, but the actual inputs of the simulator are of the corresponding high-dimensional values of $\mathbf{X}'$, as the dimensions left out are deemed unimportant.

3.4.1 Approximation properties

Here we explore some theoretical properties of the low-dimensional approximation to a simulator using the gKDR approach. For any $\mathbf{X} = \mathbf{x} \in \mathbb{R}^m$, if $\mathbf{M}(\mathbf{x})$ is known exactly, we have the eigen-decomposition (3.2). Suppose the eigenvectors and eigenvalues are partitioned as

$$\mathbf{\Lambda}(\mathbf{x}) = \begin{bmatrix} \mathbf{\Lambda}_1(\mathbf{x}) & \\ & \mathbf{\Lambda}_2(\mathbf{x}) \end{bmatrix}, \quad \mathbf{W} = [\mathbf{W}_1\ \mathbf{W}_2],$$

where $\mathbf{\Lambda}_1(\mathbf{x}) = \mathrm{diag}(\lambda_1(\mathbf{x}), ..., \lambda_d(\mathbf{x}))$ with $d < m$ consisting of the first d largest eigenvalues, $\mathbf{W}_1$ is the $m \times d$ matrix whose columns are the associated eigenvectors. Then for any $\mathbf{X}$, we can define the projected coordinates by $\mathbf{U} = \mathbf{W}_1^T\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{V} = \mathbf{W}_2^T\mathbf{X} \in \mathbb{R}^{m-d}$. Our proposed approach suggests to make inference on Y

based on $\mathbf{U}$ instead of the full explanatory variable $\mathbf{X}$. The following proposition establishes an error bound for such approximation.

Proposition 3. *For any $g \in \mathcal{H}_\mathcal{Y}$ and $\mathbf{u} \in \mathbb{R}^d$, we approximate $\mathrm{E}[g(Y)|\mathbf{X} = \mathbf{x}]$ by $\mathrm{E}[g(Y)|\mathbf{U} = \mathbf{u}]$ for any $\mathbf{x}$ such that $\mathbf{W}_1^T\mathbf{x} = \mathbf{u}$. The approximation error is bounded as follows:*

$$\|\mathrm{E}[g(Y)|\mathbf{X} = \mathbf{x}] - \mathrm{E}[g(Y)|\mathbf{U} = \mathbf{u}]\|_{L_2}^2 \leq C_1 \left(\sum_{i=d+1}^{m} b_i \lambda_i^2(\mathbf{x}) \right),$$

where C_1 is a constant depending on the domain of $\mathbf{x}$, b_i $(i = d+1, ..., m)$ are positive constants relating to $\mathbf{W}_1$ and g.

Proof. Let $G(\mathbf{x}) = \mathrm{E}[g(Y)|\mathbf{X} = \mathbf{x}]$, and $\phi_i = C_{Y\mathbf{X}} C_{\mathbf{XX}}^{-1} \frac{\partial k_{\mathcal{X}(\cdot,\mathbf{x})}}{\partial x_i} \in \mathcal{H}_\mathcal{Y}$ for $i = 1, ..., m$. Following Amini and Wainwright (2012), for any $g \in \mathcal{H}_\mathcal{Y}$, we can define a bounded linear operator $\Phi : \mathcal{H}_\mathcal{Y} \to \mathbb{R}^m$ on the Hilbert space such that $\Phi g = \begin{bmatrix} \langle \phi_1, g \rangle_{\mathcal{H}_\mathcal{Y}} & \langle \phi_2, g \rangle_{\mathcal{H}_\mathcal{Y}} & \cdots & \langle \phi_m, g \rangle_{\mathcal{H}_\mathcal{Y}} \end{bmatrix}^T$. Its adjoint is a mapping $\Phi^* : \mathbb{R}^m \to \mathcal{H}_\mathcal{Y}$, defined by the relation $\langle \Phi g, \mathbf{a} \rangle_{\mathbb{R}^m} = \langle g, \Phi^* \mathbf{a} \rangle_{\mathcal{H}_\mathcal{Y}}$ for any $g \in \mathcal{H}_\mathcal{Y}$ and $\mathbf{a} \in \mathbb{R}^m$. Therefore we have $\langle \Phi g, \mathbf{a} \rangle_{\mathbb{R}^m} = \sum_{i=1}^m a_i \langle \phi_i, g \rangle_{\mathcal{H}_\mathcal{Y}} = \langle \sum_{i=1}^m a_i \phi_i, g \rangle_{\mathcal{H}_\mathcal{Y}}$, so that for any $\mathbf{a} \in \mathbb{R}^m$, we have $\Phi^* \mathbf{a} = \sum_{i=1}^{m} a_i \phi_i$.

Defining $\mathbf{K} = \Phi\Phi^* \in \mathbb{R}^m$, it is easy to see that $\mathbf{K}_{ij} = \langle \phi_i, \phi_j \rangle_{\mathcal{H}_\mathcal{Y}}$. From the derivation of gKDR approach, the derivative of $G(\mathbf{x})$ with respect to (w.r.t) $\mathbf{x}$ is just $\nabla_\mathbf{x} G = \Phi g$ and $\mathbf{M} = \mathbf{K} = \Phi\Phi^*$. We denote the range of an operator A as $\mathrm{Ra}(A)$ and its kernel (null space) as $\mathrm{Ker}(A)$. The space $\mathrm{Ra}(\Phi^*)$ is finite-dimensional and hence closed, so we have the decomposition $\mathcal{H}_\mathcal{Y} = \mathrm{Ra}(\Phi^*) \oplus \mathrm{Ker}(\Phi)$. In particular, for any $g \in \mathcal{H}_\mathcal{Y}$, there is $\mathbf{a} \in \mathbb{R}^m$ and $g^\perp \in \mathrm{Ker}(\Phi)$ such that $g = \Phi^* \mathbf{a} + g^\perp$. Hence we obtain $\Phi g = \mathbf{M}\mathbf{a}$.

Given the projection of coordinates from $\mathbf{x}$ to $\mathbf{u}$ and $\mathbf{v}$, we can write

$$G(\mathbf{x}) = G(\mathbf{W}\mathbf{W}^T\mathbf{x}) = G(\mathbf{W}_1\mathbf{W}_1^T\mathbf{x} + \mathbf{W}_2\mathbf{W}_2^T\mathbf{x}) = G(\mathbf{W}_1\mathbf{u} + \mathbf{W}_2\mathbf{v}).$$

The gradient of G w.r.t $\mathbf{u}$ can be obtained by the chain rule as

$$\nabla_{\mathbf{u}} G = \nabla_{\mathbf{u}} G(\mathbf{W}_1\mathbf{u} + \mathbf{W}_2\mathbf{v}) = \mathbf{W}_1^T \nabla_{\mathbf{x}} G(\mathbf{x}) = \mathbf{W}_1^T \mathbf{M}\mathbf{a} = \mathbf{\Lambda}_1(\mathbf{x})\mathbf{B}^T\mathbf{a},$$

where $\mathbf{a} \in \mathbb{R}^m$ relates to g. Then it is easy to see that

$$\|\nabla_{\mathbf{u}} G\|_{L_2}^2 = \sum_{i=1}^{d} b_i \lambda_i^2(\mathbf{x}),$$

where the positive constants b_i depend on $\mathbf{W}_1$ and g, for $i = 1, ..., d$. Similarly, we have

$$\|\nabla_{\mathbf{v}} G\|_{L_2}^2 = \sum_{i=d+1}^{m} b_i \lambda_i^2(\mathbf{x}),$$

where the positive constants b_i depend on $\mathbf{W}_2$ and g, for $i = d+1, ..., m$.

We now infer $g(Y)$ based on $\mathbf{u} \in \mathbb{R}^d$ rather than $\mathbf{x} \in \mathbb{R}^m$ with $d < m$. For any $\mathbf{u}$, we have

$$\mathrm{E}[G|\mathbf{u}] = \int_{\mathbf{v}} \mathrm{E}[g(Y)|\mathbf{u}, \mathbf{v}]) \, \mathrm{d}\, P(\mathbf{v}|\mathbf{u}) = \mathrm{E}[g(Y)|\mathbf{u}].$$

Therefore for any fixed $\mathbf{u}$, we estimate $G(\mathbf{x}) = G(\mathbf{W}_1\mathbf{u} + \mathbf{W}_2\mathbf{v})$ with $\mathrm{E}[G|\mathbf{u}]$ for any $\mathbf{x} = \mathbf{W}_1\mathbf{u} + \mathbf{W}_2\mathbf{v}$, i.e. $G(\mathbf{x}) \approx \hat{G}(\mathbf{x}) = \mathrm{E}[G|\mathbf{W}_1^T\mathbf{x}] = \mathrm{E}[G|\mathbf{u}]$.

Note that for any fixed $\mathbf{u}$, $G(\mathbf{x}) = G(\mathbf{W}_1\mathbf{u} + \mathbf{W}_2\mathbf{v})$ is a function of only $\mathbf{v}$, while the approximation $\hat{G}(\mathbf{x}) = \mathrm{E}\left[G|\mathbf{u}\right]$ is in fact the average of $G(\mathbf{u}, \mathbf{v})$ over all possible $\mathbf{v}$ which is fixed. The Poincaré inequality yields

$$\|G - \hat{G}\|_{L_2}^2 \leq C_1 \|\nabla_{\mathbf{v}} G\|_{L_2}^2 = C_1 \left(\sum_{i=d+1}^{m} b_i \lambda_i^2(\mathbf{x}) \right),$$

where C_1 is a constant depending on the domain of $\mathbf{x}$. □

When $\mathbf{W}_1$ represents a sufficient dimension reduction, $\lambda_i(\mathbf{x}) = 0$ for $i = d+1, ..., m$, which implies that $\mathrm{E}[g(Y)|\mathbf{X} = \mathbf{x}] = \mathrm{E}[g(Y)|\mathbf{U} = \mathbf{W}_1^T\mathbf{x}]$ exactly. Though the result is presented with conditional mean $\mathrm{E}[g(Y)|\cdot]$ for any $g \in \mathcal{H}_{\mathcal{Y}}$, it is not limited to the first moment only. For characteristic kernels such as the

popular Gaussian RBF kernel $k(x,y) = \exp(-\|x-y\|^2/(2\sigma^2))$ and the Laplace kernel $k(x,y) = \exp(-\alpha\sum_{i=1}^m |x_i - y_i|)$, probabilities are uniquely determined by their means on the associated RKHS (Fukumizu and Leng, 2014); see also Gretton et al. (2012) for a definition of the distance between probabilities using their means.

In practice, $\mathbf{W}$ cannot be known exactly. We can only estimate a perturbed version $\tilde{\mathbf{W}} = [\tilde{\mathbf{W}}_1\ \tilde{\mathbf{W}}_2]$ instead using the eigen-decomposition of $\tilde{\mathbf{M}}_n$. Fukumizu and Leng (2014) stated that under some mild conditions, $\tilde{\mathbf{M}}_n$ converges in probability to $\mathrm{E}[\mathbf{M}(\mathbf{x})]$ with order $O_p\left(n^{-\min\{1/3,(2\beta+1)/(4\beta+4)\}}\right)$ for some $\beta > 0$. As a result, we have the following result.

Proposition 4. *For any $g \in \mathcal{H}_\mathcal{Y}$ and $\tilde{\mathbf{u}} \in \mathbb{R}^d$, we approximate $\mathrm{E}[g(Y)|\mathbf{X}=\mathbf{x}]$ by $\mathrm{E}[g(Y)|\tilde{\mathbf{U}}=\tilde{\mathbf{u}}]$ for every $\mathbf{x}$ such that $\tilde{\mathbf{W}}_1^T\mathbf{x} = \tilde{\mathbf{u}}$. As a result, we have*

$$\left\|\mathrm{E}[g(Y)|\mathbf{X}=\mathbf{x}] - \mathrm{E}[g(Y)|\tilde{\mathbf{U}}=\tilde{\mathbf{u}}]\right\|_{L_2}^2 =$$
$$O_p\left(\left(\frac{4}{\lambda_d-\lambda_{d+1}} n^{-\min\{\frac{1}{3},\frac{2\beta+1}{4\beta+4}\}}\left(\sum_{i=1}^d b_i\lambda_i^2(\mathbf{x})\right)^{\frac{1}{2}} + \left(\sum_{i=d+1}^m b_i\lambda_i^2(\mathbf{x})\right)^{\frac{1}{2}}\right)^2\right),$$

where C_1 is a constant depending on the domain of $\mathbf{x}$ and the b_i $(i=1,...,m)$ are positive constants related to $\mathbf{W}$ and g.

Proof. Denoting $\tilde{\mathbf{M}}_n = \mathrm{E}[\mathbf{M}(\mathbf{x})] + \mathbf{E}_n$ and $e_n = n^{-\min\{1/3,(2\beta+1)/(4\beta+4)\}}$, the convergence result on $\tilde{\mathbf{M}}_n$ in Fukumizu and Leng (2014) entails that for any $\epsilon > 0$, there exists a constant $C > 0$ and N_ϵ such that for any $n \geq N_\epsilon$, $P(\|\mathbf{E}_n\| < Ce_n) > 1-\epsilon$. Then there exists N'_ϵ such that for any $n \geq N'_\epsilon$, $Ce_n \leq \frac{\lambda_d-\lambda_{d+1}}{5}$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m \geq 0$ are the eigenvalues of $\mathrm{E}[\mathbf{M}(\mathbf{x})]$; where N'_ϵ can be chosen as $N'_\epsilon = \max\left\{N_\epsilon, ((\lambda_d-\lambda_{d+1})/(5C))^{-\max\{3,(4\beta+4)/(2\beta+1)\}}\right\}$.

In Golub and Van Loan (2012), the distance between subspaces that are spanned by columns of $\mathbf{W}_1$ and $\tilde{\mathbf{W}}_1$, denoted by $\mathrm{span}(\mathbf{W}_1)$ and $\mathrm{span}(\tilde{\mathbf{W}}_1)$ respectively, is defined as

$$\mathrm{dist}(\mathrm{span}(\mathbf{W}_1), \mathrm{span}(\tilde{\mathbf{W}}_1)) = \|\mathbf{W}_1\mathbf{W}_1^T - \tilde{\mathbf{W}}_1\tilde{\mathbf{W}}_1^T\| = \|\mathbf{W}_1^T\tilde{\mathbf{W}}_2\|.$$

Using Corollary 8.1.11 of Golub and Van Loan (2012), we have

$$\|\mathbf{W}_1^T\tilde{\mathbf{W}}_2\| \leq 4Ce_n/(\lambda_d - \lambda_{d+1}).$$

Therefore, $\|\mathbf{W}_1^T\tilde{\mathbf{W}}_2\| = O_p\left(4e_n/(\lambda_d - \lambda_{d+1})\right)$. We also note that $\|\mathbf{W}_2^T\tilde{\mathbf{W}}_2\| \leq \|\mathbf{W}_2\|\|\tilde{\mathbf{W}}_2\| = 1$. Then for any $\mathbf{x}$, we have the following approximation to $G(\mathbf{x})$,

$$G(\mathbf{x}) \approx \tilde{G}(\mathbf{x}) = \mathrm{E}\left[G|\tilde{\mathbf{W}}_1^T\mathbf{x}\right] = \mathrm{E}\left[G|\tilde{\mathbf{u}}\right].$$

Defining $\tilde{\mathbf{v}} = \tilde{\mathbf{W}}_2^T\mathbf{x}$ and following the same reason as Proposition 3, we have, for any fixed $\tilde{\mathbf{u}}$

$$\|G - \tilde{G}\|_{L_2}^2 \leq C_1\|\nabla_{\tilde{\mathbf{v}}}G\|_{L_2}^2,$$

where C_1 is some constant. Since $\nabla_{\tilde{\mathbf{v}}}G = \mathbf{W}_2^T\tilde{\mathbf{W}}_2\nabla_{\mathbf{v}}G + \mathbf{W}_1^T\tilde{\mathbf{W}}_2\nabla_{\mathbf{u}}G$, we have

$$\|G - \tilde{G}\|_{L_2}^2 \leq C_1\|\nabla_{\tilde{\mathbf{v}}}G\|_{L_2}^2 \leq C_1\left(\|\mathbf{W}_2^T\tilde{\mathbf{W}}_2\nabla_{\mathbf{v}}G\|_{L^2} + \|\mathbf{W}_1^T\tilde{\mathbf{W}}_2\nabla_{\mathbf{u}}G\|_{L_2}\right)^2.$$

The result holds by plugging in the respective terms. □

The approximation procedure using the dimension reduction generates an "innovative simulator" $\tilde{f}$ on the reduced input space of $\mathbf{U} = \tilde{\mathbf{W}}_1^T\mathbf{X}$, which is however not deterministic. Assuming there are two distinct inputs $\mathbf{X}_1$ and $\mathbf{X}_2$ with the respective outputs $Y_1 \neq Y_2$, it may happen that $\tilde{\mathbf{W}}_1^T\mathbf{X}_1 = \tilde{\mathbf{W}}_1^T\mathbf{X}_2$, i.e. the approximated simulator $\tilde{f}$ may yield different outputs given the same input. The low-dimensional stochastic simulator $\tilde{f}$ can nevertheless be emulated, for example using a GP with nugget effect assuming the effect of the dropped components is relatively small and simple. The overall approximate error of the final emulator $\hat{f}$ to f can be decomposed into $\|\hat{f} - f\| \leq \|f - \tilde{f}\| + \|\tilde{f} - \hat{f}\|$, where the first term in the right hand side is due to the low-dimensional approximation which has been investigated in Proposition 4, and the second term depends on the emulation procedure.

3.4.2 Choice of parameters and structural dimension

When applying the proposed framework for emulation, several parameters need to be specified properly. For example, the parameters in the kernels and the regularisation parameter ϵ_n. The cross validation approach can be used for tuning such parameters as in many nonparametric statistical methods. In addition, it is also required to choose an appropriate structural dimension d to construct an accurate emulator.

One of the possible ways is to decide or estimate d within the dimension reduction procedure. Fukumizu and Leng (2014) pointed out that it might not be practical to select d based on asymptotic analysis of some test statistics, as in many existing dimension reduction techniques, when the dimension is high and the sample size is small. They mentioned that the ratio of the sum of the largest d eigenvalues over the sum of all the eigenvalues, $\sum_{i=1}^{d} \lambda_i / \sum_{i=1}^{m} \lambda_i$, might be useful in identifying the conditional independence of Y and $\mathbf{X}$ given $\mathbf{B}^T\mathbf{X}$. In addition, Proposition 4 shows that the approximation error decreases as a function of $\lambda_d - \lambda_{d+1}$. As discussed in Constantine and Gleich (2015), d might be chosen such that $\lambda_d - \lambda_{d+1}$ is maximised. However, we may notice that the approximation error also depends on the squares of the eigenvalues with some unknown weights b_i. Therefore it seems to be not very practical to select d based on the eigenvalues only.

On the other hand, Fukumizu and Leng (2014) suggested to select d based on the following analysis rather than the dimension reduction procedure when dimension reduction serves as a pre-processing step. For example, the ultimate goal of our proposed framework here is to construct an accurate emulator, hence it is intuitive to select the structural dimension that produces the best predictive performance. Therefore, in the following numerical studies, we select d as well as other parameters for the gKDR approach using simple trial-and-error or more formal cross validation approach based on the predictive accuracy of the respective emulators.

3.5 Numerical simulations

In this section, we conduct two numerical studies. In the first study, the proposed emulation framework using the gKDR approach is compared with several alternatives of dimension reduction methods and the full emulation on a PDE problem. This problem set up allows the computation of gradients explicitly. In the second study, we illustrate the emulation framework with an application to tsunami modelling where AS cannot be applied because of the lack of explicit gradients; we also provide a comparison to other dimension reduction methods. Throughout the simulations, the GPML code using maximum likelihood method implemented by Rasmussen and Williams (2006) is employed for the emulation assuming a linear form mean function with intercept, and a squared exponential correlation function.

3.5.1 Study 1: elliptic PDE with explicit gradients available

In this example, we investigate the elliptic PDE problem with random coefficients as studied in Constantine et al. (2014). Let $u = u(\mathbf{s}, \mathbf{x})$ satisfy the linear elliptic PDE

$$-\nabla_{\mathbf{s}} \cdot (a \nabla_{\mathbf{s}} u) = 1, \;\; \mathbf{s} \in [0,1]^2.$$

The homogeneous Dirichlet boundary conditions are set on the left, top and bottom boundary (denoted by Γ_1) of the spatial domain of $\mathbf{s}$, and a homogeneous Neumann boundary condition is imposed on the right side of the spatial domain denoted Γ_2. The coefficients $a = a(\mathbf{s}, \mathbf{x})$ are modelled by a truncated Karhunen-Loeve (KL) type expansion $\log(a(\mathbf{s}, \mathbf{x})) = \sum_{i=1}^{m} x_i \gamma_i \phi_i(\mathbf{s})$, where the x_i are i.i.d. standard Normal random variables, and $\phi_i(\mathbf{s}), \gamma_i$ are the eigenpairs of the correlation operator $C(\mathbf{s}, \mathbf{t}) = \exp(\beta^{-1} \|\mathbf{s} - \mathbf{t}\|_1)$.

The target value is a linear function of the solution, $f(\mathbf{x}) = \int_{\Gamma_2} u(\mathbf{s}, \mathbf{x}) / |\Gamma_2| d\mathbf{s}$. The problem is discretised using finite element method on a triangulation mesh, then f and $\nabla_{\mathbf{x}} f$ can be computed as a forward and adjoint problem; see Constantine et al. (2014) for more details. We choose $m = 100$ and examine two cases of the correlation lengths $\beta = 1$ or $\beta = 0.01$. Therefore the original input space is $\mathcal{X} = \mathbb{R}^{100}$ with standard Normal distribution and the output $f(\mathbf{x})$ is univariate.

The gKDR approach is applied to reduce the dimension of the problem using M samples. We also compare with several popular alternative dimension reduction techniques: AS (here possible due to the explicit gradients), SIR, SIR-II (Li, 1991), sliced average variance estimation (SAVE) (Cook and Weisberg, 1991), MAVE and PFC. After reducing the dimension of the problem, the GP emulator is trained using a Latin Hypercube design of $10d$ points on the reduced d-dimensional space so that the whole procedure needs $M + 10d$ samples in total using each dimension reduction method. For comparison, we also emulate the problem on the original 100-dimensional input space directly with $M + 10d$ samples which is the full emulation. The gKDR approach is implemented in Matlab by K. Fukumizu which is available on his homepage http://www.ism.ac.jp/~fukumizu/. The Matlab code for AS and solving the PDE by Constantine et al. (2014) is available on https://bitbucket.org/paulcon/active-subspace-methods-in-theory-and-practice. For SIR, SAVE and PFC, the codes are provided in the Matlab LDR-package (https://sites.google.com/site/lilianaforzani/ldr-package), and SIR-II is implemented by simply modifying the SIR code. For MAVE, the Matlab code by Y. Xia is available from http://www.stat.nus.edu.sg/~staxyc/. The associated parameters in some methods, such as the kernel and regularisation parameters for gKDR, the number of slices for the sliced methods and the degree of polynomial basis for PFC, are chosen in a simple trial-and-error way by trying several values and selecting the best.

The final emulators are used to make prediction on a testing set of n evaluations $\{f_1, ..., f_n\}$ that are differ from the training set, where $f_i = f(\mathbf{x}_i)$ and $\mathbf{x}_i \in \mathbb{R}^{100}$ is drawn randomly from multivariate standard Normal distribution. The predictive performance is measured by the normalised predictive root-mean-square-error (PRMSE)

$$\text{Normalised PRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - \hat{f}_i)^2}}{\max_i f_i - \min_i f_i},$$

where $\hat{f}_i$ is the prediction (predictive mean) using emulation. The associated computing time is also recorded with three parts: T1 for running the simulator, T2 for

estimating the dimension reduction and T3 for training the emulator and making prediction. Note that T1 includes the time devoted to run the simulator M times when using all the dimension reduction methods. It also includes the time used to compute the gradients for AS and the additional $10d$ runs for the full emulation. T2 is zero for the full emulation since there is no dimension reduction procedure. T3 also includes the time for running the simulator $10d$ times on the designed points except the full emulation.

In this study, we choose $M = 300$ and $d = 1, ..., 5$. Table 3.1 presents the results on a testing set with $n = 500$ evaluations using different emulation approaches and Figure 3.2 shows an example of the associated computing time when $\beta = 1$ and $d = 5$. Compared with the full emulation results, by reducing the dimension properly, the predictive accuracy can be improved, especially when the correlation length is long ($\beta = 1$). Also, as a result, the computing time for training GP emulator (T3) decreases dramatically. In terms of predictive accuracy, AS naturally performs the best, as it is using exact gradients, followed by gKDR. MAVE, SIR and PFC are better than SIR-II and SAVE, but PFC does not work very well when $\beta = 1$ and MAVE spends more computing time on dimension reduction. Most methods yield smaller errors for $\beta = 1$ than $\beta = 0.01$, except PFC and full emulation. Unlike the other techniques, AS employs exact gradients which might explain its advantage. However, as shown in Figure 3.2, the computing time T1 for AS is about two orders of magnitude longer than the others making the method most computationally expensive. Computing gradients $\nabla_{\mathbf{x}} f$ is sometimes impossible, e.g. for the tsunami simulation in the next study. This restricts the applicability of AS method to a few applications. To summarise, when the exact gradients are computable, the proposed gKDR approach is able to produce comparable results (though not as good) as the AS method that uses exact gradients, and outperforms the other SDR methods in most cases. However the computational cost of applying gKDR approach is much less than for the AS approach. In fact, gKDR not only is able to find the SDR accurately and efficiently, but also can be applied in a wide range of scenarios where complicated variable types or very high dimensions are involved. The next applica-

tion into tsunami simulations provides a snapshot of its wide capability when there are few applicable alternatives.

Table 3.1: Normalised PRMSEs at 500 testing sites using emulation on the full input space (Full) or combined with different dimension reduction techniques.

$\beta = 1$								
d	gKDR	AS	SIR	SIR-II	SAVE	MAVE	PFC	Full
1	0.116	0.126	0.125	0.153	0.153	0.126	0.152	0.097
2	0.044	0.007	0.025	0.153	0.153	0.020	0.140	0.095
3	0.032	0.011	0.024	0.152	0.152	0.019	0.120	0.095
4	0.024	0.012	0.024	0.150	0.150	0.024	0.080	0.093
5	0.024	0.011	0.026	0.150	0.150	0.083	0.071	0.092
$\beta = 0.01$								
1	0.037	0.033	0.043	0.169	0.169	0.039	0.161	0.032
2	0.033	0.028	0.039	0.169	0.169	0.038	0.160	0.032
3	0.033	0.029	0.039	0.167	0.167	0.039	0.034	0.032
4	0.033	0.025	0.039	0.167	0.167	0.039	0.033	0.032
5	0.033	0.024	0.038	0.167	0.167	0.037	0.033	0.032

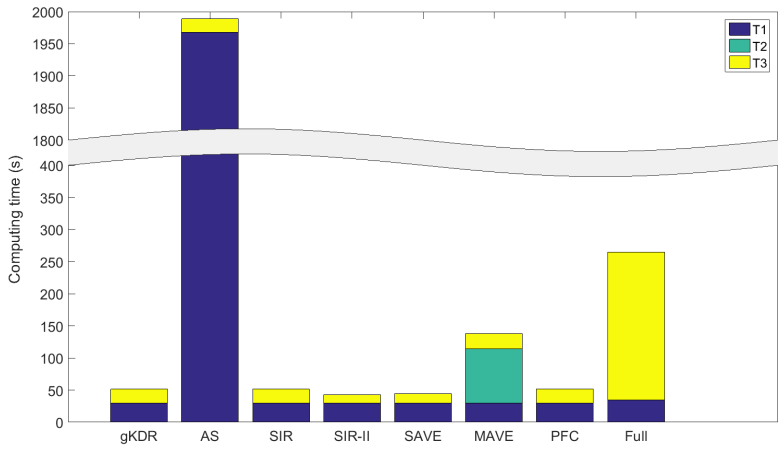


Figure 3.2: Computing time (in seconds) for the emulation using different approaches when $\beta = 1$ and $d = 5$.

3.5.2 Study 2: tsunami emulation where no gradients available

Here we apply the proposed general framework to investigate the impact of the uncertainties in the bathymetry on tsunami modelling, where the bathymetry is included as a high-dimensional input.

A synthetic bathymetry surface is created in the (s_1, s_2) coordinate system to

conduct tsunami simulations as shown in Figure 3.3 (a). For simplicity, we assume that the seabed elevation only vary along the first coordinate s_1. Though simple, it still captures the typical continental characteristics: the continental shelf spans from shore line ($s_1 = 0$) to around $s_1 = -25$ km at the water depth of around 150 m; the continental slope is between $s_1 = -25$ km and $s_1 = -75$ km with water depth of $150 \sim 1500$ m; then it is the deep ocean.

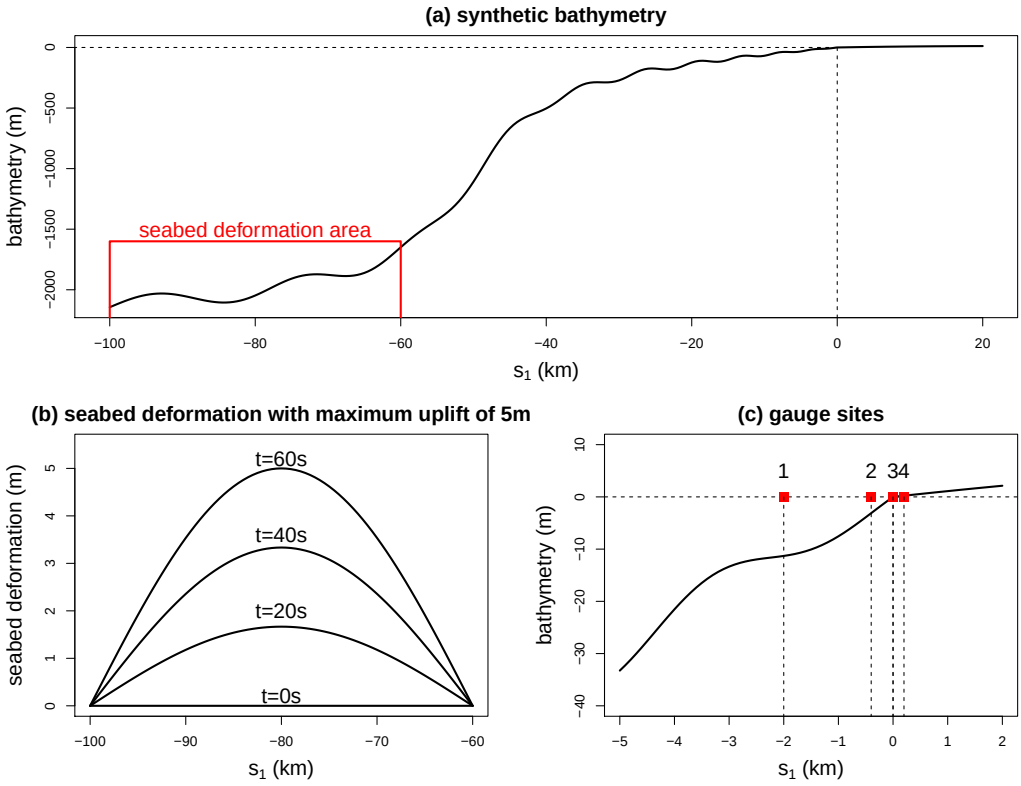


Figure 3.3: (a) Synthetic bathymetry; (b) seabed uplift when $h_{max} = 5$ m; (c) gauge sites.

To mimic the realistic boat tracks of oceanic surveys for bathymetry data collection, some irregular lines are drawn. We consider two levels of survey density which are denoted by survey level 1 and 2 respectively. Considering that the surveys are usually constrained within budgets, the total lengths of the two level surveys are fixed at 1000 and 100 km. To account for different possible survey traces, 20 samples of boat tracks are drawn at each level of density; see in Figure 3.4 three samples per level for illustration. In this study, we only consider the impact of the uncertainties in the bathymetry within the area $(s_1, s_2) \in [-40000, 0] \times [-5000, 5000]$ as shown with a blue rectangle in Figure 3.4. The bathymetry at other locations are

fixed at the true values. This assumption is based on the physical knowledge that deep ocean has a relatively small influence on tsunami waves.

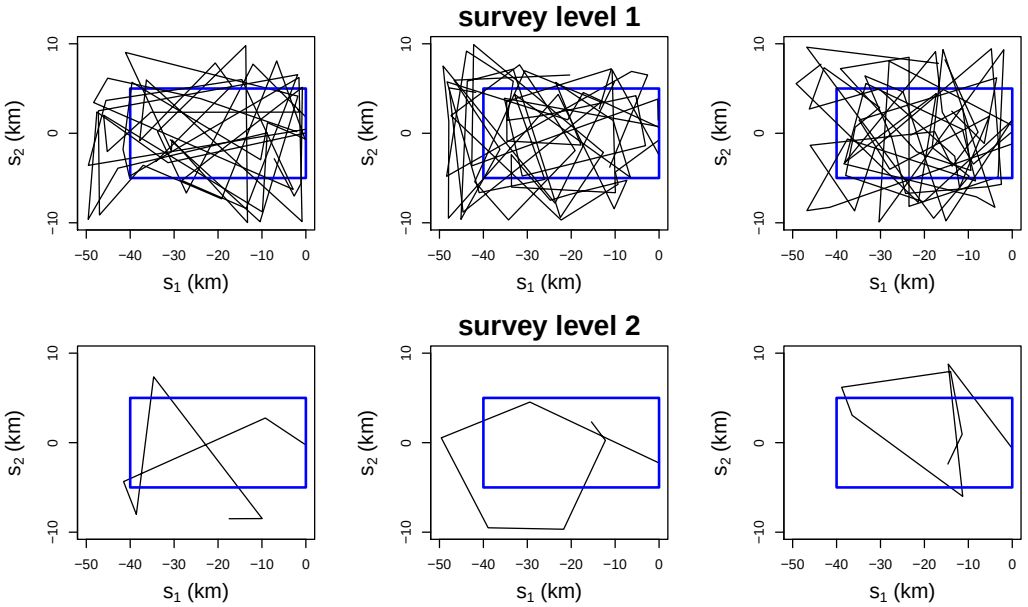


Figure 3.4: Three samples of boat tracks at two levels of survey density; the bathymetry within the blue rectangle are assumed uncertain.

Along the possible boat tracks, observations of bathymetry are collected every 500 m. Then the whole bathymetry surface can be modelled using the SPDE approach and inferred using the INLA method. Given observations of bathymetry $\mathbf{z} = (z_1, ..., z_n)'$ at locations $\mathbf{s} = (\mathbf{s}_1, ..., \mathbf{s}_n)'$, it is assumed that $z_i = Z(\mathbf{s}_i) + \epsilon_i$, $i = 1, ..., n$, where the unknown bathymetry surface $Z(\mathbf{s})$ is Gaussian field with Matérn covariance function. Lindgren et al. (2011) noted that $Z(\mathbf{s})$ also satisfies the SPDE $\tau(\kappa^2 - \Delta)^{\alpha/2} Z(\mathbf{s}) = W(s)$, where τ and κ relate to the variance and correlation length of the Matérn covariance. With a finite elements representation $Z(\mathbf{s}) = \sum_{k=1}^{m} w_k \psi_k(\mathbf{s})$ over an appropriate triangular mesh, a stochastic weak solution to the SPDE can be approximated. It is shown that the coefficients $\mathbf{w} = (w_1, ..., w_m)^T$ can be approximated by a Gaussian Markov random field, i.e. $\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}^{-1})$ for $\mathbf{Q}$ is sparse. Note that bivariate splines could be used (Liu et al., 2015) to reduce the number of parameters required for specific approximation order, which is good but not enough for dimension reduction. Then we build

the following hierarchical spatial model,

$$\begin{aligned}\mathbf{z}|\mathbf{w},\boldsymbol{\theta} &\sim \mathrm{N}(\mathbf{A}\mathbf{w},\sigma_e^2\mathbf{I}),\\ \mathbf{w}|\boldsymbol{\theta} &\sim \mathrm{N}(\mathbf{0},\mathbf{Q}(\boldsymbol{\theta})^{-1}),\\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}),\end{aligned}$$

where $\mathbf{A}_{ij} = \psi_j(\mathbf{s}_i)$ and $\boldsymbol{\theta}$ contains all the hyperparameters. Since $\mathbf{w}$ uniquely determines the bathymetry, it is the *de facto* input for the uncertain bathymetry. In this study, we build a mesh for the finite elements representation in the SPDE approach as shown in Figure 3.5 (a). The dense triangles in the middle cover the uncertain bathymetry area and the outer extension is to avoid boundary effect. There are 3200 nodes that influence the bathymetry, hence the uncertain input for bathymetry is of dimension 3200. Given each boat track and the associated observations $\mathbf{z}$, 20 samples of the finite element coefficients are drawn from the posterior $\pi(\mathbf{w}|\mathbf{z})$ to construct the possible initial bathymetry. Thus there are 400 sets of possible initial bathymetry in total at each survey level. Figure 3.6 shows their sample mean and standard deviation.

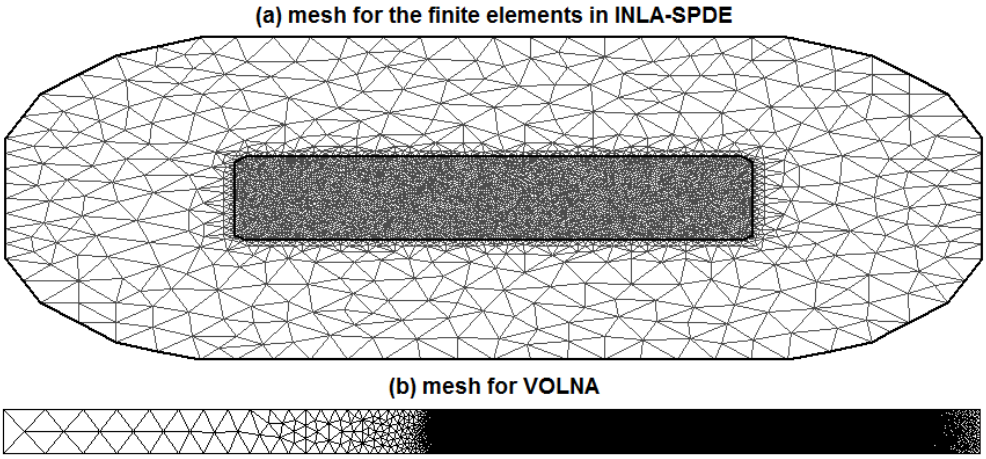


Figure 3.5: (a) Mesh for the SPDE approach; (b) mesh for VOLNA.

Tsunami waves are triggered by the following simplified seabed deformation,

$$\mathrm{d}\, z(s_1, s_2; t) = \frac{t}{60} \cdot h_{max} \cdot \sin\left(\frac{s_1 + 100000}{-60000 + 100000}\pi\right) \cdot I_{\{-100000 \le s_1 \le -60000, 0 \le t \le 60\}},$$

where $\mathrm{d}\, z(s_1, s_2; t)$ is the seabed uplift at location $\mathbf{s} = (s_1, s_2)$ and time t, h_{max} denotes the maximum seabed uplift; see Figure 3.3 (b) for example. We take 5 different values $h_{max} = 1, ..., 5$ m. These values are evenly combined with the uncertain initial bathymetry. Thus there are two sources of uncertainties: $\mathbf{w}$ for bathymetry and h_{max} for tsunami source, where $\mathbf{w}$ is high-dimensional.

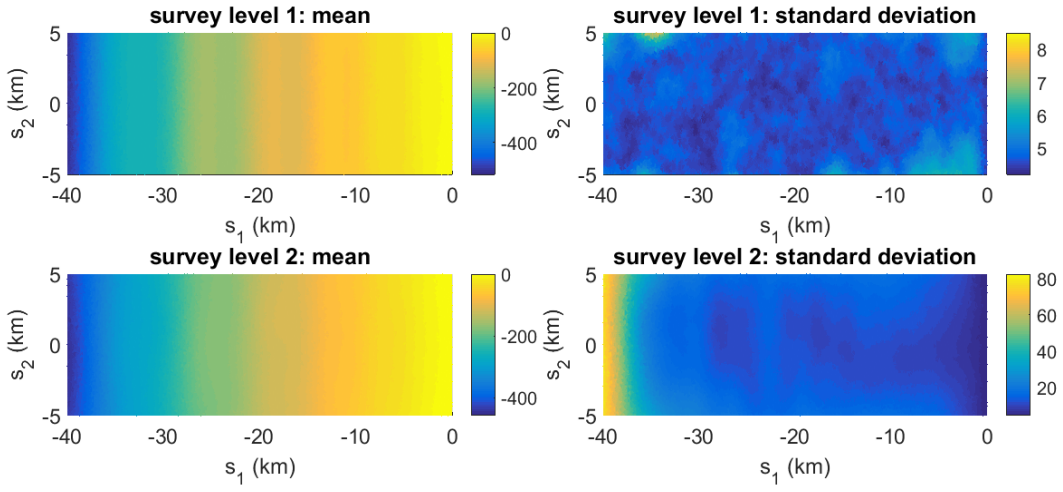


Figure 3.6: Sample mean and standard deviation of the bathymetry input; note the different scales of standard deviation for survey level 1 and 2.

We employ the tsunami code VOLNA (Dutykh et al., 2011), an advanced nonlinear shallow water equation solver using the finite volume method on a high performance computing facility, to simulate the tsunami events. The computational domain and mesh for VOLNA are presented in Figure 3.5 (b). There are 120,661 triangles and 61,068 nodes in the mesh, where the coarse triangles in both ends are added to avoid boundary reflection. The output of the simulation is chosen to be $\Delta\eta(\mathbf{s}) = \max \eta_t(\mathbf{s}) - \eta_0(\mathbf{s})$, where $\eta_t(\mathbf{s})$ is the free surface elevation at simulation time t and location $\mathbf{s}$. $\Delta\eta$ represents the maximum wave height at off shore locations or the maximum inundation depth at on shore locations. For illustration, we consider simulation values at gauge 1: $(-2000, 0)$, gauge 2: $(-400, 0)$, gauge 3: $(0, 0)$ and gauge 4: $(200, 0)$, which are at far shore, near shore, shore line and land respectively; see Figure 3.3 (c).

The simulation results are presented in Figure 3.7, along with those using the true bathymetry as shown in red lines and those using the sample mean bathymetry

as shown in green dash lines. We can see that $\Delta\eta$ increases with h_{max} but also shows variation due to the uncertain inputs $\mathbf{w}$ for fixed h_{max}, especially at gauges 2-4 around the shore line. In general, the simulations with sample mean bathymetry would deviate from true values, while those with random bathymetry samples can cover the true events quite well. The survey level also has significant influence but shows different scales at different gauges. The wider range of possible simulation values with coarser survey level 2 indicates that the uncertainty in the bathymetry would spread the tsunami waves out to simulate more extreme scenarios and such effect could be amplified around the shore line.

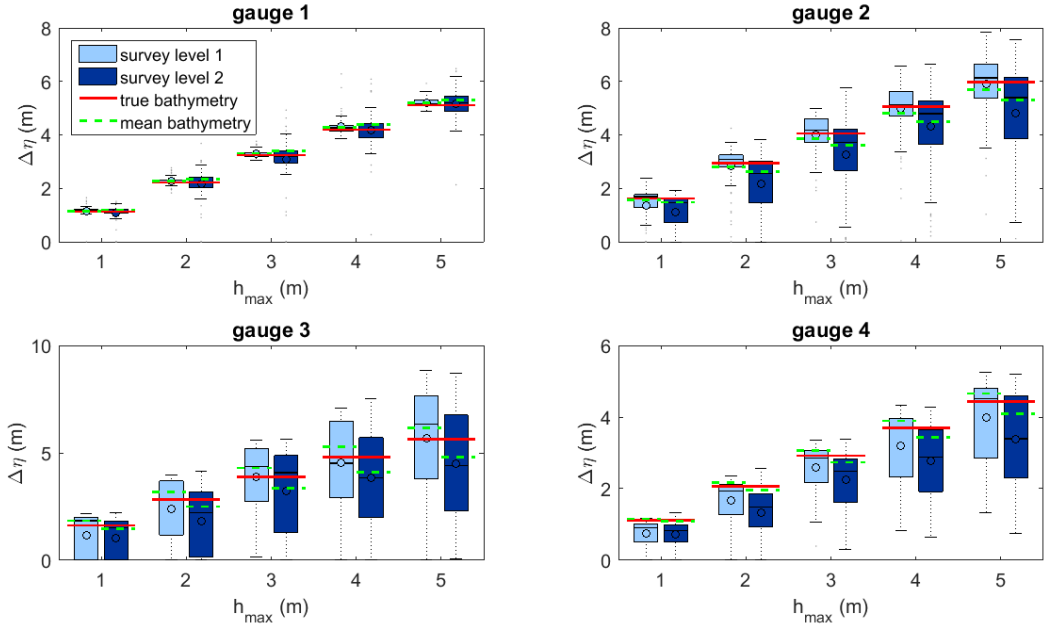


Figure 3.7: Simulation values with different inputs $(\mathbf{w}, h_{max})$ at four gauges.

Following the procedure in Section 3.4, we can construct a low-dimensional emulator for such high dimensional simulator with 3200 input parameters for the bathymetry ($\mathbf{w}$) and 1 parameter for the seabed deformation (h_{max}). Denoting the VOLNA code with f, the output can be represented as $\Delta\eta = f(\mathbf{w}, h_{max})$. Because Figure 3.7 displays a significant relationship between h_{max} and $\Delta\eta$, we keep it as a separate input in the emulator and reduce the dimension of $\mathbf{w}$ only. In this case, we try to find a projection matrix $\mathbf{B}$ such that $\mathbf{w} \perp (\Delta\eta, h_{max})|\mathbf{B}^T\mathbf{w}$. The conditional independence just implies the sufficiency of $\mathbf{B}^T\mathbf{w}$, i.e. $p(\Delta\eta|h_{max}, \mathbf{w}) =$

$\tilde{p}(\Delta\eta|h_{max}, \mathbf{B}^T\mathbf{w})$ (Yin and Hilafu, 2015). Therefore $(\Delta\eta, h_{max})$ is regarded as a temporary output when applying gKDR to reduce the dimension of $\mathbf{w}$.

For the gKDR approach, Gaussian RBF kernels $k(x, y) = \exp(\|x - y\|^2/(2\sigma^2))$ are deployed for both $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ but with different parameters σ^2. Following Fukumizu and Leng (2014), we have the parameterisation $\sigma_{\mathcal{X}} = c_1\sigma_{med}(\mathbf{X})$, $\sigma_Y = c_2\sigma_{med}(\mathbf{Y})$ where $\sigma_{med}(\cdot)$ is the median of pairwise distance of the data. The regularization parameter ϵ_n is fixed at 10^{-5}. There are three parameters to be specified properly: c_1, c_2, and d, the number of directions included in the emulator. We consider possible candidates $c_1, c_2 \in [0.5, 1, 5, 10, 15, 20]$, $d \in [1, ..., 5]$ here. Then, based on each of the possible parameters combination, we can construct a GP emulator on the low-dimensional inputs $(h_{max}, \mathbf{B}^T\mathbf{w})$ and make predictions on the new inputs $(\tilde{h}_{max}, \mathbf{B}^T\tilde{\mathbf{w}})$.

For comparison, we also apply alternative dimension reduction techniques to construct the low-dimensional approximations. Due to the complexity of the VOLNA code and multi-system nature of the whole simulation set up, the gradients of simulation values with respect to the inputs are not computable. Hence the active subspace method cannot be employed. Most of the methods in Study 1 cannot be applied directly because of the need for partial dimension reduction, or the "large p, small n" feature, i.e. there are much more input parameters than the number of simulations. We consider two extensions to PFC and SIR. The partial PFC (PPFC) method following Kim (2011) is implemented based on the R package ldr (Adragni and Raim, 2014) to find the reduction on $\mathbf{w}$ only meanwhile taking the effect of h_{max} into account. Note that PFC is not developed for the problem where $p > n$ or $p \gg n$. Another method we compare to is the sequential sufficient dimension reduction (SSDR) by Yin and Hilafu (2015). It is specifically proposed to overcome the "large p, small n" difficulty by decomposing the variables into pieces each of which has $p_1 < n$ variables so that conventional dimension reduction methods can be applied. The projective resampling approach (Li et al., 2008) with SIR is employed. The R code for SSDR by Yin and Hilafu (2015) is available from http://wileyonlinelibrary.com/journal/rss-datasets.

To measure the predictive performance and select proper parameters, a 10-fold cross validation is employed. For each survey level, 400 simulations are divided evenly into 10 groups. Each group is retained as testing set once, while the other nine groups are used to estimate the projection matrix $\mathbf{B}$ using gKDR, PPFC or SSDR approaches and train the respective GP emulator. Table 3.2 presents the normalised PRMSEs from the cross validation for each survey level and gauge using different dimension reduction techniques and the respective effective dimension selected. It is shown that the effective dimension is around 1-4 in most cases except for gauge 1 at survey level 2 using SSDR which is 8. For the survey level 1, the selected dimension is generally smaller than that for the survey level 2. This is consistent with the observation that the uncertainty scale in survey level 2 is higher and more complex which makes the dimension reduction more difficult. The errors of survey level 1 are in general smaller than those of survey level 2. This implies that as the uncertainties in the bathymetry increase, it gets more difficult to make accurate predictions using emulation. The methods gKDR and SSDR outperform PPFC in all cases, especially in survey level 1 where the normalised PRMSEs can be 50% lower in some cases. In survey level 1, the errors of the gKDR approach are slightly larger than those of SSDR for gauge 2-4 where the normalised PRMSEs using SSDR are around $1.1\% \sim 3.7\%$ lower. But in survey level 2, the gKDR approach is more accurate than the SSDR approach for all gauges with reduction of normalised PRMSEs at 1.0% for gauge 1 and $10.1\% \sim 18.9\%$ for gauge 2-4. Therefore, gKDR is comparable with SSDR in survey level 1 but works much better than SSDR in survey level 2 when there are more uncertainties involved. We can conclude that the proposed GP emulation framework combined with the gKDR dimension reduction approach is effective and accurate for this complicated tsunami simulator and overall it outperforms the alternatives.

To investigate the impact of the training set size on the predictive performance of the proposed emulation framework with gKDR approach, we conduct repeated random sub-sampling cross validations with various training set sizes. We consider training set size as 2%,5%,10%,20%,...,90% and test set size as 10% of the total

Table 3.2: Normalised PRMSEs of the 10-fold cross validation using GP emulation and different dimension reduction methods with the effective dimension selected in the parentheses.

gauge	survey level 1			survey level 2		
	gKDR	PPFC	SSDR	gKDR	PPFC	SSDR
1	0.031(2)	0.078(1)	0.033(1)	0.095(4)	0.096(2)	0.096(8)
2	0.099(2)	0.138(1)	0.097(1)	0.134(3)	0.175(3)	0.149(4)
3	0.091(2)	0.187(1)	0.090(1)	0.129(2)	0.210(3)	0.159(4)
4	0.082(2)	0.144(2)	0.079(1)	0.106(3)	0.141(3)	0.121(3)

400 simulations. For each training set size, the sampling is repeated 50 times. The parameters c_1, c_2, d are fixed at those values selected through the above 10-fold cross validation. Figure 3.8 displays the normalised PRMSEs with various training set sizes. In general the predictive errors decrease as the training set size increases, and eventually converge to a relatively flat level.

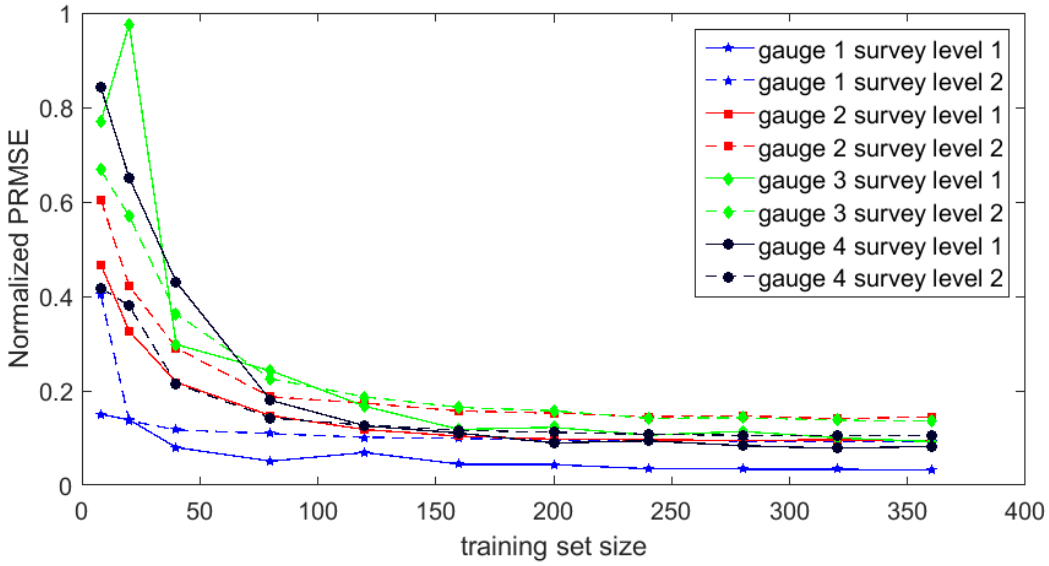


Figure 3.8: Normalised PRMSEs with various training set sizes.

In the end, we apply the resulting emulator to predict the simulation values over a large number of new inputs. The predictions can be used for probabilistic risk assessment and many other purposes. For illustration, 10000 samples of $(\tilde{h}_{max}, \tilde{\mathbf{w}})$ are drawn where $\tilde{h}_{max}$ are drawn from Normal distribution $\mathrm{N}(3, 1)$ truncated at 0 and 5. For each survey level, 100 samples of possible boat tracks are drawn randomly. Given the observations along each boat track, 100 samples of $\tilde{\mathbf{w}}$ are drawn

from the posterior. In most of the current tsunami research work, the bathymetry is usually considered as fixed, which would neglect the possible uncertainty in the outputs. For comparison, we conduct another set of simulations with the 5 possible values of h_{max}, but with a fixed $\mathbf{w}$ taken to be the sample mean as shown in Figure 3.6. In this case, h_{max} is the only uncertain parameter. Then we can also make another set of predictions on the 10000 samples of $\tilde{h}_{max}$ only. The predictions for these two cases are presented in Figure 3.9. We can see that at gauge 1 it makes no significant difference to include the uncertainty in the bathymetry or not, as the impact is relatively small on the far shore waves as shown in Figure 3.7. However the impact of the uncertainty in the bathymetry on the simulation values is more significant at gauges around the shore line. The distributions are shifted, skewed and spread out, covering more extreme events with larger $\Delta\eta$. These features are potentially important, for example in the catastrophe models that are widely used in (re)insurance to calculate the possible losses. We will discuss more details about the use of tsunami hazard assessment with catastrophe models in Appendix C. This simulation example is also used to illustrate the significant impact of the uncertainties in the bathymetry on financial losses.

3.6 Discussion

We proposed a joint framework for emulation of high-dimensional simulators with dimension reduction. The gKDR approach is employed to construct low-dimensional approximations to the simulators. The approximations retain most of the information about the input-output behaviour and make the emulation much more efficient. Both theoretical properties and numerical studies have demonstrated the efficiency and accuracy of the proposed approach and its advantages over other dimension reduction techniques. Our method can be applied for many purposes of uncertainty quantification such as risk assessment, sensitivity analysis and calibration, with great perspectives in real world applications. There are some practical issues when applying the proposed framework. The hyperparameters in gKDR and the number of dimensions to be included in the emulator need to be specified prop-

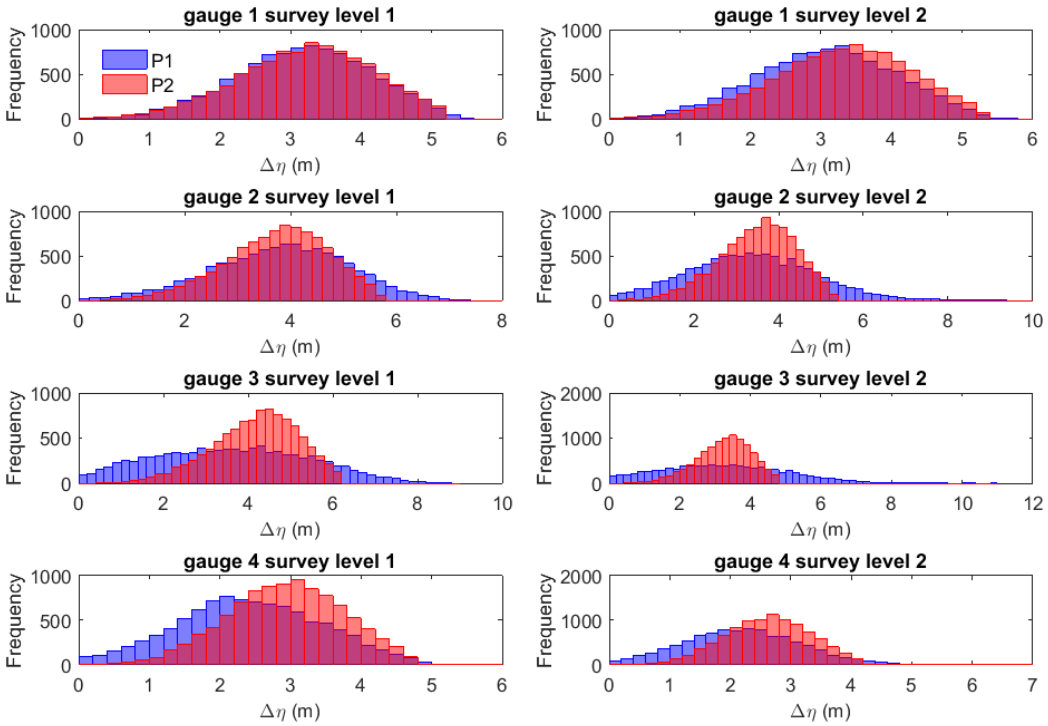


Figure 3.9: Histogram of predictions of 10000 events of uncertain seabed uplift when taking into account the uncertainties in the bathymetry (P1) or not (P2).

erly. In practice, a simple trial and error procedure could be applied, especially when the results are not very sensitive to the choices. The cross validation steps could also benefit from parallel computing technique. The sample size also affect the predictive ability of the final emulator, as sufficient samples are needed to estimate the dimension reduction accurately. A diagnostic plot of predictive errors with increasing number of sizes such as Figure 3.8 could help identify the convergence. After determining the dimension reduction, a sufficient number of training samples with a proper design are often required to train the emulator in order to balance the computational cost and accuracy. The benefits of our approach are multiple. One can tackle uncertainty quantification tasks for complex models where boundary conditions are of high dimension. Beyond tsunami modelling, in climate, weather or geophysical sciences, uncertainty quantification studies would become tractable and potentially offer solutions to important scientific problems.

Chapter 4

Joint Modelling of Multiple Spatial Surveys

4.1 Introduction

In geostatistics, a continuous spatial process of the quantity of interest is usually studied with a finite set of measurements. For example, bathymetry and topography data are collected and used in geo-mapping and subsequently earthquake or tsunami modelling; disease prevalence surveys are conducted to study the spatial variation and evolution of diseases. In most of these surveys, data are usually sampled highly irregularly over space. Moreover, there are often several available surveys over the same region. It is clearly helpful to extract information of the common spatial process from multiple available surveys.

The idea of combining multiple surveys has been investigated using methods in meta analysis and small area statistics (Elliott and Davis, 2005, Lohr and Rao, 2006, Manzi et al., 2011, Moriarity and Scheuren, 2001, Turner et al., 2009). More recently, Giorgi et al. (2015) proposed a multivariate generalized linear geostatistical model for multiple prevalence surveys to accommodate both spatial and temporal variations. Their combined model is able to improve the spatial inference in terms of root-mean-square-error, and coverage of nominal 95% confidence intervals, by considering the heterogeneity among surveys in a joint model. Most of these studies focus on prevalence surveys that are usually different from geoscientific surveys.

For example, the method proposed by Giorgi et al. (2015) assumes that at least one of the surveys provides an unbiased representation of the spatial process. This assumption is probably not reasonable for bathymetric and topographic surveys. It is indeed difficult to find a single survey or even a group of surveys with similar properties that can be considered as a well-measured unbiased "gold standard". In addition, the conventional inference tools employed in prevalence surveys could be too computationally intensive, or even prohibitive, for the extraordinarily large data found in geosciences.

In bathymetric data processing, the most frequently used method is gridding with continuous curvature splines in tension, encoded in the open source software Generic Mapping Tools (GMT) (Wessel and Smith, 1998). It has been used by NOAA to produce Digital Elevation Models (DEMs). The problem of multiple surveys in the construction of bathymetric data products have already been addressed. Hell and Jakobsson (2011) enhanced the functionality of GMT to handle heterogeneous bathymetric data sets. They took into consideration the local data density in the interpolation: regions that are covered with dense data are gridded at higher resolutions than those with sparse data. Then by stacking the grids at multiple resolutions, final digital bathymetric models (DBMs) were shown to be superior to remove-restore grids (Becker et al., 2009) and splines in tension grids (Smith and Wessel, 1990). Hell and Jakobsson (2011) only considered the differentiation in data densities among the data sets. But the differences in other aspects such as spatial coverage and accuracy are not discussed, and thus would yield potential improvement if acknowledged in the modelling process.

In this chapter, we use the SPDE geostatistical approach (Lindgren et al., 2011) to model the whole surface of bathymetry. The SPDE approach provides efficient inference with INLA (Rue et al., 2009), so that it is applicable to large spatial data sets. We consider a joint hierarchical latent Gaussian model for multiple surveys based on the SPDE approach. Each survey is modelled separately on top of the common latent Gaussian field. Thus different characteristics in the surveys, e.g. measurement accuracy and density, can be treated. Moreover, due to some con-

straints in budget or safety, or specific scientific purposes, bathymetric surveys usually display preferential sampling features, that is to say the measured locations may be dependent on the values of the underlying spatial process (Diggle et al., 2010). For example, shallow region may be avoided during the surveys because of the danger of stranding. The spatial locations themselves may provide information about the underlying spatial process and the preferential sampling feature can be included in the joint model to improve the inference. One of the advantages of geostatistical approaches over the interpolation techniques such as GMT is the probabilistic treatment to the prediction that could be used in the associated uncertainty quantification to produce more reliable results; see Chapter 3 for an example. Though the main motivation and focus in this chapter are bathymetric surveys, the model could be potentially generalised or adjusted to accommodate other types of spatial surveys.

The plan of this chapter is as follows. In Section 4.2, we describe the joint model for the combined and simultaneous analysis of multiple spatial surveys. The joint model is hierarchical and includes separate layers for the surveys. The latent field is modelled with the SPDE approach and the preferential sampling patterns are also included as log-Gaussian Cox processes. In Section 4.3, we illustrate the joint model and compare it with other models on Matérn fields under different scenarios. We also demonstrate an application to a set of realistic bathymetric surveys. Section 4.4 contains some discussion.

4.2 Joint model with the SPDE approach

Suppose there are m surveys for a spatial process $f(\mathbf{s})$ over a region $D \subset \mathbb{R}^2$, denoted by $\mathbf{Y}_1, ..., \mathbf{Y}_m$. For each $i = 1, ..., m$, $\mathbf{Y}_i = (y_{i1}, ..., y_{in_i})^T$ contains n_i observations of the spatial process $f(\mathbf{s})$ at locations $\mathbf{S}_i = (\mathbf{s}_{i1}, ..., \mathbf{s}_{in_i})^T$. For illustration, we assume that $y_{ij} = f(\mathbf{s}_{ij}) + \epsilon_{ij}$, where $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_i^2)$, for $j = 1, ..., n_i$, $i = 1, ..., m$. This implies that the observational errors of different surveys are characterised by separate distributions of different variances to account for their respective measurement accuracy. The additive Normal assumption may be adjusted accordingly in the joint model for different applications. When the surveys might

be preferentially conducted, the point pattern of each survey is modelled as a inhomogeneous Poisson process with intensity $\lambda_i(\mathbf{s}) = \exp\{a_i + b_i f(\mathbf{s})\}$ following Diggle et al. (2010). The parameter a_i relates to the baseline of the sampling density when there is no preferential sampling ($b_i = 0$). The relationship between the sampling preference and the value of spatial process is determined by b_i where $b_i > 0$ implies that locations with higher values are more likely to be sampled while $b_i < 0$ indicates preference to locations of lower values. Therefore, conditioned on the underlying spatial process $f(\mathbf{s})$, we have the following joint model for the multiple surveys with preferential sampling,

$$\begin{aligned} y_{ij}|\mathbf{s}_{ij}, f &\sim \mathrm{N}\left(f(\mathbf{s}_{ij}), \sigma_i^2\right) \\ \mathbf{s}_{ij}|f &\sim \text{Poisson}\left(\exp\{a_i + b_i f(\mathbf{s}_{ij})\}\right), \end{aligned} \tag{4.1}$$

for $i = 1, ..., m$ and $j = 1, ..., n_i$.

As proposed in Lindgren et al. (2011), the spatial process $f(\mathbf{s})$ where $\mathbf{s} \in \mathbb{R}^2$ can be modelled as a zero-mean Gaussian field with Matérn covariance function

$$c(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa h)^\nu K_\nu(\kappa h),$$

where h is the Euclidean distance, K_ν is the modified Bessel function of the second kind and order $\nu > 0$, $\kappa > 0$ controls the nominal correlation range and σ^2 is the marginal variance. The integer value of ν determines the mean-square differentiability of the underlying process. As noted in Lindgren et al. (2011), the Matérn field $f(\mathbf{s})$ can be seen as a solution to the stochastic partial differential equation (SPDE)

$$\tau(\kappa^2 - \nabla)^{\alpha/2} f(\mathbf{s}) = W(\mathbf{s}),$$

where $\alpha = \nu - 1$, $\nabla = \frac{\partial^2}{\partial s_1^2} + \frac{\partial^2}{\partial s_2^2}$ is the Laplacian operator, τ controls the marginal variance through the relationship

$$\tau^2 = \frac{\Gamma(\nu)}{\Gamma(\nu+1)4\pi\kappa^{2\nu}\sigma^2},$$

and $W(\mathbf{s})$ is spatial white noise process. The Matérn field is represented with piecewise linear finite elements as $f(\mathbf{s}) = \sum_{k=1}^{n} w_k \phi_k(\mathbf{s})$, where $\mathbf{w} = (w_1, ..., w_n)^T$ are multivariate Gaussian weights and $\{\phi_k(\mathbf{s})\}_{k=1}^{n}$ is a set of piecewise linear basis functions. By solving the SPDE, it has been shown in Lindgren et al. (2011) that the Gaussian weights $\mathbf{w}$ can be approximated with a Gaussian Markov random field, i.e. $\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}^{-1})$, where the precision matrix $\mathbf{Q}$ is sparse.

Under the Gaussian assumption about $f(\mathbf{s})$, the spatial point pattern $\mathbf{S}_i$ for each survey is in fact a log-Gaussian Cox Process (Diggle et al., 2010), where the Poisson intensity function is modelled as a log-Gaussian field. The likelihood of the inhomogeneous Poisson process $\mathbf{S}_i$ is

$$\pi(\mathbf{S}_i | f(\mathbf{s})) = \exp\left\{ |D| - \int_D \exp\left\{a_i + b_i f(\mathbf{s})\right\} d\mathbf{s} \right\} \prod_{\mathbf{s}_{ij} \in \mathbf{S}_i} \exp\left\{a_i + b_i f(\mathbf{s}_{ij})\right\}.$$

Because of the integration of the unknown intensity function $\lambda_i(\mathbf{s})$ over the whole domain D, the likelihood is analytically intractable. Simpson et al. (2016) proposed to numerically approximate this integral when using the SPDE model for $f(\mathbf{s})$. By employing a deterministic integration rule $\int_D g(\mathbf{s}) d\mathbf{s} \approx \sum_{l=1}^{p} \tilde{\alpha}_l g(\tilde{\mathbf{s}}_l)$ with fixed integration nodes $\{\tilde{\mathbf{s}}_l\}_{l=1}^{p}$ and the associated weights $\{\tilde{\alpha}_l\}_{l=1}^{p}$, the log-likelihood of $\mathbf{S}_i$ can be approximated as follows,

$$\begin{aligned}
\log\{\pi(\mathbf{S}_i | f)\} &= |D| - \int_D \exp\left\{a_i + b_i f(\mathbf{s})\right\} d\mathbf{s} + \sum_{j=1}^{n_i} \left\{a_i + b_i f(\mathbf{s}_{ij})\right\} \\
&\approx |D| - \sum_{l=1}^{p} \tilde{\alpha}_l \exp\left\{ a_i + b_i \sum_{k=1}^{n} w_k \phi_k(\tilde{\mathbf{s}}_l) \right\} + \\
&\quad \sum_{j=1}^{n_j} \left\{ a_i + b_i \sum_{k=1}^{n} w_k \phi_k(\mathbf{s}_{ij}) \right\} \\
&\propto \tilde{\boldsymbol{\alpha}}^T \exp(a_i + b_i \mathbf{B}\mathbf{w}) + \mathbf{1}^T (a_i + b_i \mathbf{A}_i \mathbf{w}),
\end{aligned} \tag{4.2}$$

where $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, ..., \tilde{\alpha}_p)^T$, $\mathbf{B}$ is a $p \times n$ matrix whose (l,k)-th element is $[\mathbf{B}]_{lk} = \phi_k(\tilde{\mathbf{s}}_l)$, and $\mathbf{A}_i$ is a $n_i \times n$ matrix with $[\mathbf{A}_i]_{jk} = \phi_k(\tilde{\mathbf{s}}_{ij})$. By writing $\log \boldsymbol{\eta}_i = (a_i + b_i \mathbf{w}^T \mathbf{B}^T, a_i + b_i \mathbf{w}^T \mathbf{A}_i^T)^T$, $\boldsymbol{\alpha}_i = (\tilde{\boldsymbol{\alpha}}^T, \mathbf{0}_{n_i \times 1}^T)^T$ and constructing some pseudo-

observations $\mathbf{Z}_i = (\mathbf{0}^T_{p\times 1}, \mathbf{1}^T_{n_i\times 1})^T$ where 1's or 0's represent the associated locations are observed or not corresponding to $\{\tilde{\mathbf{s}}_k\}_{k=1}^p$ and $\{\mathbf{s}_{ij}\}_{j=1}^{n_i}$ respectively, the approximate likelihood (4.2) is

$$\pi(\mathbf{S}_i|f) \approx \pi(\mathbf{Z}_i|f) \propto \prod_{j=1}^{n_i+p} (\alpha_{ij}\eta_{ij})^{Z_{ij}} \exp(-\alpha_{ij}\eta_{ij}),$$

which is the core of the likelihood of $N + p$ independent Poisson random variables with intensity parameters $\alpha_{ij}\eta_{ij}$ as noticed in Simpson et al. (2016).

To complete the integration scheme, one has to make choices for the integration nodes $\{\tilde{\mathbf{s}}_k\}_{k=1}^p$ and the associated weights $\{\tilde{\alpha}_k\}_{k=1}^p$. It is intuitive to take advantage of the triangulation structure of the SPDE approach. For example, Simpson et al. (2016) proposed to set $\tilde{\mathbf{s}}_k$ at the triangulation nodes and attach to each node a region V_k for which the value of the basis function $\phi_k(\mathbf{s})$ is greater than the value of all the other basis functions. This leads to a dual mesh of the triangulation that can be constructed by joining the centroids of the triangles. By setting $\tilde{\alpha}_k = |\mathbf{V}_k|$, $k = 1, ..., n$, this integral approximation is of second-order accuracy on a regular grid and first-order accuracy on an irregular mesh. As mentioned in Simpson et al. (2016), the integration scheme can be constructed in other ways such as applying the optimal Gaussian integration rule on each triangle. Various choices for integration nodes and weights can be found in the literature of finite element methods or numerical analysis.

By applying the integration rule as discussed above, the joint model (4.1) can be approximated using the SPDE approach as follows,

$$\begin{aligned}
\mathbf{Y}_i|\mathbf{S}_i, \mathbf{w}, \boldsymbol{\theta} &\sim \mathrm{N}\left(\mathbf{A}_i\mathbf{w}, \sigma_i^2\mathbf{I}_{n_i}\right), \quad i = 1, ..., m \\
Z_{ij}|\mathbf{w}, \boldsymbol{\theta} &\sim \mathrm{Poisson}\left(\alpha_{ij}\eta_{ij}\right), \quad j = 1, ..., n_i + n \\
\mathbf{w}|\boldsymbol{\theta} &\sim \mathrm{N}\left(\mathbf{0}, \mathbf{Q}^{-1}(\boldsymbol{\theta})\right) \\
\boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}),
\end{aligned} \tag{4.3}$$

where Z_{ij}, α_{ij} and η_{ij} have been defined previously, $\boldsymbol{\theta}$ contains the hyperparameters including σ_i^2, a_i, b_i and those in the SPDE approach such as κ and τ. The approxi-

mate joint model (4.3) is a latent Gaussian model and can be fitted using the INLA approach for efficient Bayesian inference. Note that the first layer of the hierarchical model is not necessarily Normal, but can be any other sensible likelihoods to accommodate the specific data features. Additional covariates may be included as well to improve the modelling if applicable. When some of the surveys do not display significant pattern of preferential sampling, the second layer could be removed accordingly to simplify the model.

4.3 Numerical experiments

In this section, we examine and illustrate the proposed hierarchical joint model with two numerical studies, one of which is for some given theoretical Matérn fields and the other is for the purpose of realistic bathymetry mapping. All the analysis are run using the INLA package in R.

4.3.1 Study 1: synthetic Matérn fields

In this study, the proposed model is applied to make inference for the latent Matérn field with mean zero and covariance $C(h) = \sigma^2\{2^{\nu-1}\Gamma(\nu)\}^{-1}(\kappa h)^\nu K_\nu(\kappa h)$, where $\nu = 1$ is fixed. Given a realisation of the latent field f_0, two surveys are conducted at locations $\mathbf{S}_1$ and $\mathbf{S}_2$ respectively, which are drawn from inhomogeneous Poisson point processes with intensities $\boldsymbol{\lambda}_1 = \exp\{a_1 + b_1 f_0\}$ and $\boldsymbol{\lambda}_2 = \exp\{a_2 + b_2 f_0\}$. The respective responses are assumed to follow Gaussian distributions with different variances, i.e. $\mathbf{Y}_1 \sim \mathrm{N}\left(\beta_0 + f_0(\mathbf{S}_1), \sigma_1^2\right)$ and $\mathbf{Y}_2 \sim \mathrm{N}\left(\beta_0 + f_0(\mathbf{S}_2), \sigma_2^2\right)$.

Figure 4.1 presents a realisation of the Matérn latent Gaussian field with $\sigma^2 = 3$ and $\kappa = 5$. Then we consider two surveys following the Poisson sampling intensities with $a_1 = 1$, $a_2 = 1.5$, $b_1 = 0.5$ and $b_2 = 1.0$ as shown in Figure 4.1. Both surveys tend to cover more higher values, while $\mathbf{S}_1$ has less observations and is less preferential to high values than $\mathbf{S}_2$. For the rest of the simulation set up, we let $\beta_0 = 10$, $\sigma_1 = 0.5$ and $\sigma_2 = 0.05$, which means that survey 1 is less accurate than survey 2. Then we are able to make inference on the latent field with these surveys combined. Employing the SPDE approach with the mesh and its associated dual mesh as shown in Figure 4.2, two different models given f_0 as follows are

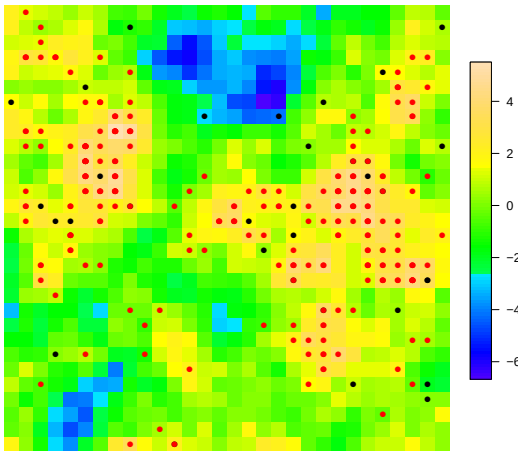


Figure 4.1: Sampling locations of survey 1 ($\mathbf{S}_1$, black) and survey 2 ($\mathbf{S}_2$, red) based on a realisation of a Matérn latent Gaussian field as shown in the background.

compared.

M1: Letting $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T)^T$ and $\mathbf{S} = (\mathbf{S}_1^T, \mathbf{S}_2^T)^T$, we assume $\mathbf{Y}|f_0 \sim \mathrm{N}(f_0(\mathbf{S}), \sigma_0^2)$ and $\mathbf{S}$ completely random.

M2: For $i = 1, 2$, we assume that $\mathbf{Y}_i|f_0 \sim \mathrm{N}(f_0(\mathbf{S}_i), \sigma_i^2)$ and $\mathbf{S}_i \sim \mathrm{Poisson}(\exp\{a_i + b_i f_0(\mathbf{S}_i)\})$.

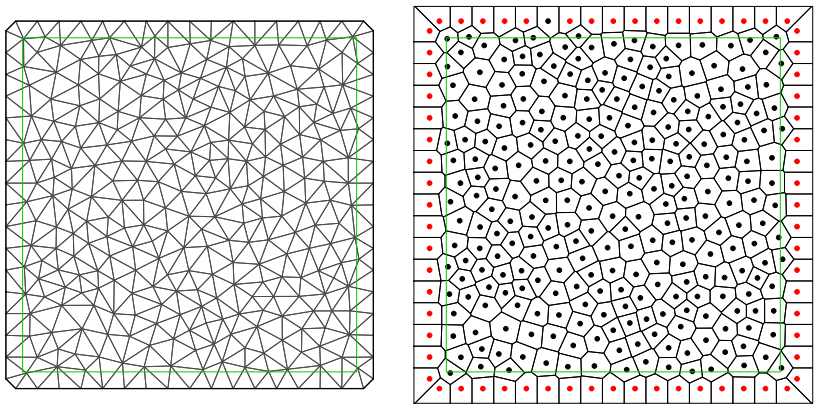


Figure 4.2: Mesh (left) for the SPDE approach and the dual mesh (right) to define the integral scheme.

M1 is a simple model that combines the two surveys together and fits a single ordinary spatial model. M2 involves separate treatments with preferential sampling feature for the surveys. Both models fit naturally into a Bayesian hierarchical modelling and can be inferred using the SPDE approach with the INLA efficiently. Fig-

ure 4.3 presents the posterior marginals of various parameters relating to the latent field with comparison to their true values. The posterior distributions of β_0, κ and nominal range obtained from M2 are able to cover the respective true values better than M1. Nominal range is determined by κ through $\rho = \sqrt{8\nu}/\kappa$ corresponding to correlations near 0.1 at the Euclidean distance ρ. The coverage of the posterior distributions of σ^2 of M1 and M2 look similarly while M2 shows relatively larger variance. However, both models tend to overestimate σ^2. As noted in Gelman et al. (2006), this is inevitable due to the asymmetry in its parameter space with variance parameters restricted to be positive. They recommended to use the half-Cauchy prior with scale set to a reasonable value to alleviate the tendency of overestimation. As commonly agreed, the choice of prior could be influential to the Bayesian inference, but this is beyond the scope of this thesis and we stick to the default set up for priors in the INLA package.

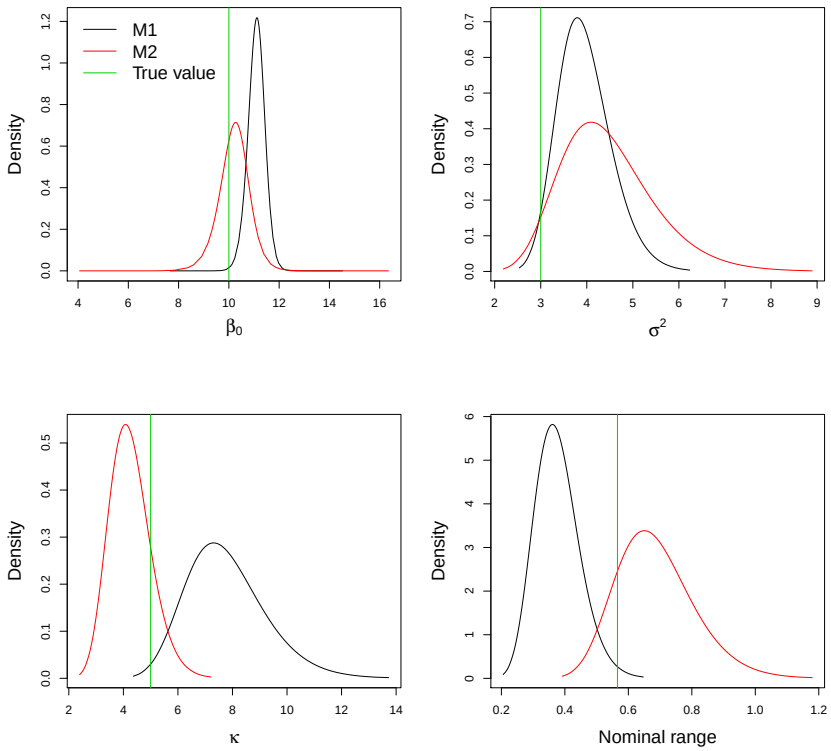


Figure 4.3: Posterior marginals of some parameters of the Matérn field using M1 and M2 based on the two surveys.

It is also able to examine how well M2 captures the preferential characteristics.

Figure 4.4 presents the posterior marginals of the parameters a_i and b_i $(i = 1, 2)$ in the Poisson intensities. It is shown that M2 is able to estimate a_1 quite well, but overestimates a_2 and underestimate b_1, b_2. Nevertheless, the main features are captured that the a_2 and b_2 are estimated generally larger than a_1 and b_1 respectively, which is consistent with the true simulation set up.

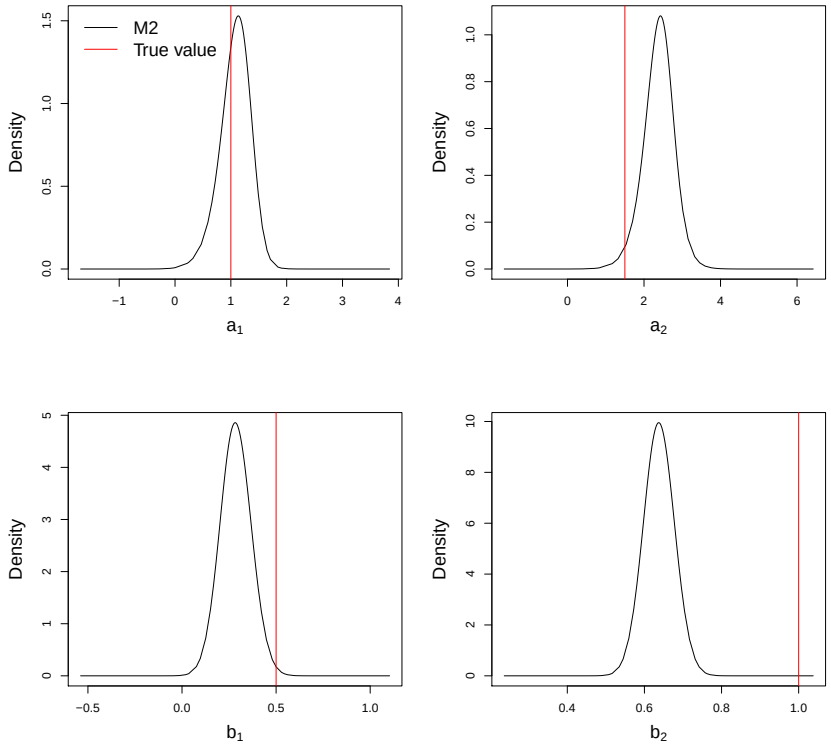


Figure 4.4: Posterior marginals of parameters for the preferential sampling using M2.

To evaluate the predictive performance of the models, we predict the latent field f_0 over a 30×30 regular gird using the two models. As in the previous chapters, the posterior means of f_0 at the grid points are taken to be the predictions. The predictive errors (prediction minus true value) are presented in Figure 4.5. It is clear that M2 yields significantly smaller errors than M1. The posterior means and standard deviations for the latent field using M2 are presented in Figure 4.6.

In the example above, we have only considered one realisation of a Matérn field with two specific surveys. This simulation set up is not representative of various situations. To assess the predictive performance of the models more widely and

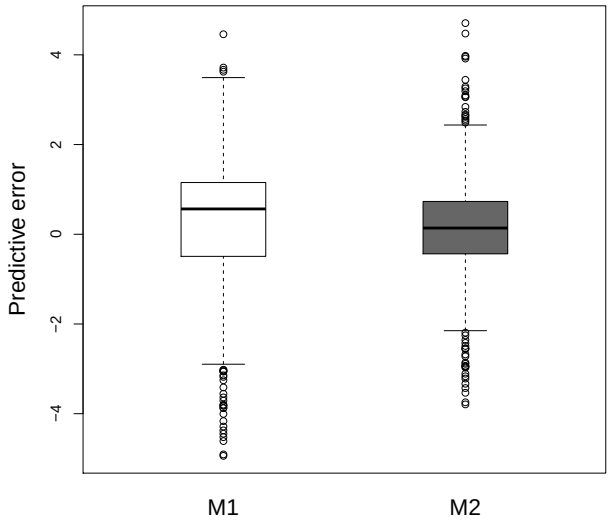


Figure 4.5: Boxplot of the predictive errors of models M1 and M2.

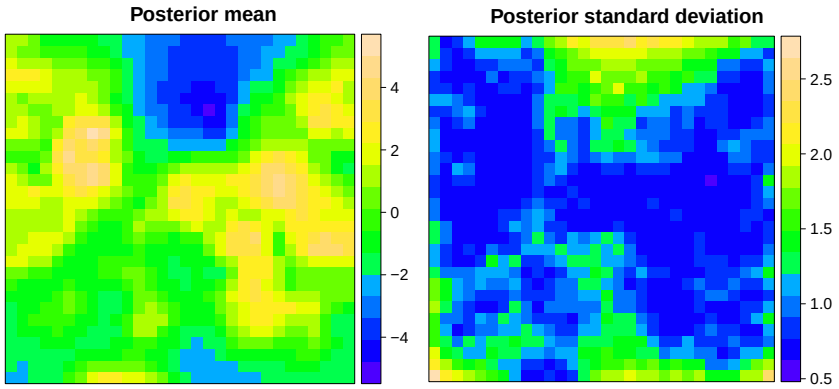


Figure 4.6: Posterior mean and standard deviation for the latent field using M2.

investigate the sensitivity of the models to the Matérn field parameters and sampling characteristics, different scenarios are now studied. We consider the following four different latent Gaussian Matérn fields (LGMFs): LGMF 1 with $\sigma^2 = 1$ and $\kappa = 5$, LGMF 2 with $\sigma^2 = 3$ and $\kappa = 5$, LGMF 3 with $\sigma^2 = 1$ and $\kappa = 2$, and LGMF 4 with $\sigma^2 = 3$ and $\kappa = 2$. Figure 4.7 shows a sample realisation of each field. We can see that the variability in the field elevations increases as σ^2 increases, and the spatial variability increases as κ increases. For each LGMF, 500 realisations of the latent field are drawn randomly.

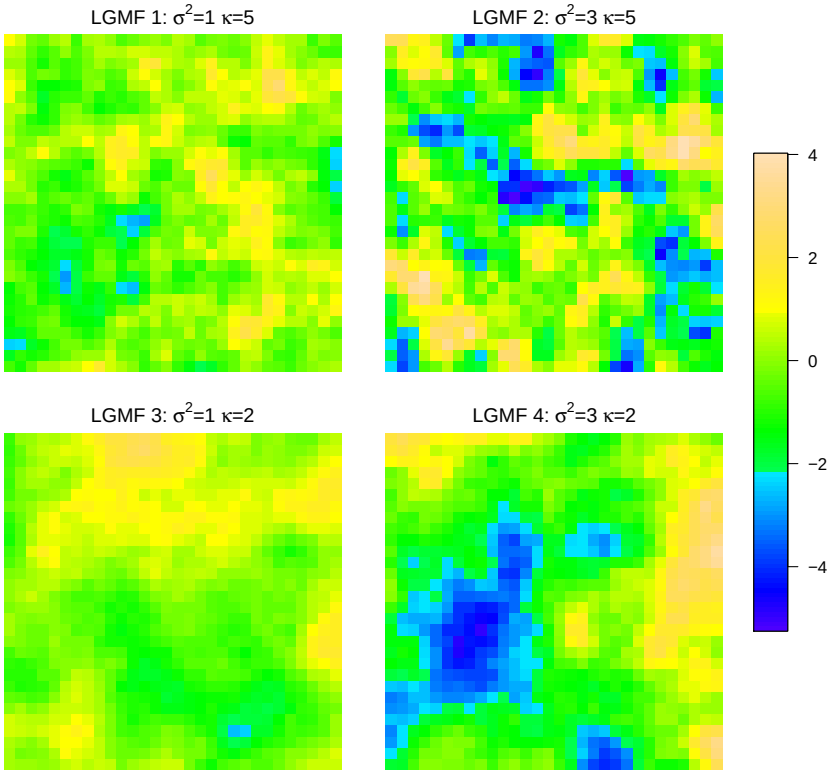


Figure 4.7: Sample realisations of four different latent Gaussian Matérn fields.

Based on each realisation, the impact of different cases of survey strategies of two types are explored. In the survey strategy type 1, both of the two surveys are more likely to be conducted at locations where the elevations are high, i.e. $b_1 > 0$ and $b_2 > 0$. But in the survey strategy type 2, one of the surveys prefers high values while the other prefers low values, say $b_1 > 0$ and $b_2 < 0$. Furthermore, different values for the parameters a_i and b_i $(i = 1, 2)$ in each survey strategy type are considered. Table 4.1 presents the different cases of surveys with the associated

parameters. We can see that Case 1-6 are of survey strategy type 1 where both two surveys in each case prefer high values while Case 7-12 are of survey strategy type 2 where the two surveys in each case have opposite preference. In each case, since we are more interested in the impact of the preferentiality, a_1 and a_2 are set to the same value. But we still consider two values for $a_1 = a_2$ to see the influence on the prediction. In survey strategy type 1, b_1 is fixed at 0.5 while b_2 is varied at 0.5, 0.8 or 1.2 so that different relative levels of preferentiality can be compared. In survey strategy type 2, when $b_1 = 0.5$ we let $b_2 = -0.5$ (Case 7/10) or $b_2 = -1.0$ (Case 8/11) to see the impact of different levels of preferentiality. We also consider the case $b_1 = -0.5$ and $b_2 = 1.0$ (Case 9/12) where the preferentiality is reversed as Case 8/11. The nugget terms for the two surveys are kept the same as in the previous example say $\sigma_1 = 0.5$ and $\sigma_2 = 0.05$.

Table 4.1: Parameters in the Poisson intensity of different cases of survey strategy.

Survey strategy type 1						
Parameters	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
a_1	1.0	1.0	1.0	1.5	1.5	1.5
b_1	0.5	0.5	0.5	0.5	0.5	0.5
a_2	1.0	1.0	1.0	1.5	1.5	1.5
b_2	0.5	0.8	1.2	0.5	0.8	1.2
Survey strategy type 2						
Parameters	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12
a_1	1.0	1.0	1.0	1.5	1.5	1.5
b_1	0.5	0.5	-0.5	0.5	0.5	-0.5
a_2	1.0	1.0	1.0	1.5	1.5	1.5
b_2	-0.5	-1.0	1.0	-0.5	1.0	-1.0

We generate 500 realisations of each Matérn field and the associated two surveys in each situation. The prediction of the latent field is carried out using M1 and M2 respectively based on the data each time. The average normalised PRMSEs (ANPRMSEs) over the 500 repetitions is used as a measure of the predictive performance. For comparison, the relative percentage increase (RPI) of the ANPRMSEs by using simple model M1 instead of M2 is computed as

$$\mathrm{RPI} = 100 \times \frac{\mathrm{ANPRMSE_{M1}} - \mathrm{ANPRMSE_{M2}}}{\mathrm{ANPRMSE_{M2}}}.$$

Table 4.2: Relative percentage increase (RPI) in the normalised PRMSEs using M1 instead of M2 based on different LGMFs and cases of survey strategy.

Survey strategy type 1						
LGMF	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
1	4.18	5.80	9.55	4.35	6.59	10.96
2	8.90	12.57	18.04	9.40	13.23	18.80
3	5.86	8.90	10.04	5.70	7.04	9.45
4	7.10	11.67	15.05	6.10	9.44	11.92
Survey strategy type 2						
LGMF	Case 7	Case 8	Case 9	Case 10	Case 11	Case 12
1	−0.72	0.82	1.57	−0.22	2.45	2.52
2	1.27	8.20	7.30	1.60	9.19	8.35
3	1.62	2.86	4.10	3.21	4.23	4.51
4	2.11	5.36	4.14	2.05	5.03	4.99

Table 4.2 presents the RPIs for different cases and LGMFs. In survey strategy type 1, we can see that by using M1 mistakenly instead of M2, the errors can be increased by up to nearly 19%. When the difference between the two surveys increases, say from Case 1 to 3 or Case 4 to 6, the improvement of using M2 is more significant. The RPIs of LGMF 2 and 4 are larger than those of LGMF 1 and 3. This implies that as the variability of the underlying fields increase, it gets more beneficial to use M2 instead of M1. The effect of a_i $(i = 1, 2)$ seems to be dependent on the LGMFs. For LGMFs 1 and 2 where $\kappa = 5$, RPIs become slightly larger when a_i increases (from Case 1-3 to Case 4-6). But for LGMFs 3 and 4 where $\kappa = 2$, RPIs become smaller when a_i increases. In survey strategy type 2, the benefit of using M2 instead of M1 is generally smaller than that in survey strategy type 1. The two surveys in each case of type 2 have opposite preferentiality making the observations somehow evenly distributed over both high and low values so that the effect of preferential sampling is alleviated and the advantage of M2 over M1 is weakened. Nevertheless, M2 still outperforms M1 in most of the cases. Similar features are observed, e.g. as the difference between the two surveys or the variability in the LGMFs increases, M2 is getting more advantageous than M1. For Case 7 and 10, M2 performs slightly worse than M1. In the two cases, the latent field is relatively simple, $b_1 = 0.5$ and $b_2 = -0.5$ which makes the overall observations cover both high and low values evenly. Hence the simple model M1 is sufficient while the more

complicated model M2 makes the inference less accurate without enough data.

In the example above, the nugget terms are fixed across different scenarios. Hence it is difficult to distinguish between the effect of preferential sampling features and the effect of separate modelling of different surveys in terms of the benefit of using M2. We now investigate the impact of different levels of σ_i on the prediction, with $\mathbf{S}_i$ completely random sampled to exclude the effect of preferential sampling for $i = 1, 2$. Table 4.3 presents the different cases of nugget terms. Without loss of generality, it is assumed that survey 1 is more accurate. In addition, we consider the number of observations in the two surveys are equal and take different values $n_1 = n_2 = 25, 50, 100, 200$. The following two models are compared.

M1: Letting $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T)^T$ and $\mathbf{S} = (\mathbf{S}_1^T, \mathbf{S}_2^T)^T$, we assume $\mathbf{Y}|f_0 \sim \mathrm{N}(f_0(\mathbf{S}), \sigma_0^2)$ and $\mathbf{S}$ completely random.

M2: For $i = 1, 2$, we assume that $\mathbf{Y}_i|f_0 \sim \mathrm{N}(f_0(\mathbf{S}_i), \sigma_i^2)$ and $\mathbf{S}_i$ completely random.

Table 4.3: Different cases of nugget terms in the observations.

Nugget	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9
σ_1	0.05	0.05	0.05	0.10	0.10	0.10	0.20	0.20	0.20
σ_2	0.10	0.20	0.40	0.20	0.40	0.80	0.40	0.80	1.60

Table 4.4 presents the relative percentage increase in the average normalised PRMSEs over 500 Monte Carlo repetitions using M1 instead of M2 in different situations. It is shown that, unlike Table 4.2, there are more negative entries implying that the simple model M1 outperforms the proposed model M2 in the associated situations, especially when the number of data is small say 25 or 50. In such case, though M2 is representing the true model, there are not enough data to achieve an accurate inference so that it is less effective than M1. We also notice that as the difference between the two surveys increases, e.g. from Case 1/4/7 to 3/6/9, the performance of M2 is getting better. Such improvement of using M2 seems to be amplified by the increasing number of data (n_1, n_2) or increasing noise level (σ_1, σ_2). Though M2 is less effective in some situations, the values of RPIs range around $-1.80\% \sim -0.01\%$, which means that M2 is still able to yield comparable

Table 4.4: Relative percentage increase (RPI) in the normalised PRMSEs using M1 instead of M2 based on different LGMFs and cases of nugget terms.

LGMF	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9
$n_1 = n_2 = 25$									
1	-0.11	-0.04	-0.01	-0.19	-0.21	0.76	-0.31	0.27	3.06
2	-0.10	-0.30	-0.20	-0.17	-0.19	-0.21	-0.23	-0.37	1.06
3	-0.05	0.05	0.77	-0.09	0.60	3.92	-0.45	2.67	10.08
4	-0.03	-0.01	0.16	-0.07	0.14	1.18	-0.15	0.59	4.66
$n_1 = n_2 = 50$									
1	-0.27	-0.61	-0.32	-0.61	-0.35	2.32	-0.72	1.79	10.97
2	-0.53	-0.75	-0.81	-0.73	-0.77	-0.26	-0.96	-0.44	3.43
3	-0.37	-0.17	1.62	-0.40	1.25	8.06	0.21	6.12	21.32
4	-0.22	-0.34	0.02	-0.40	-0.09	2.43	-0.44	1.84	10.04
$n_1 = n_2 = 100$									
1	-1.20	-1.22	-0.81	-0.73	-0.28	3.42	-0.15	3.31	14.77
2	-1.80	-1.67	-1.64	-1.13	-1.14	-0.42	-0.65	0.14	5.11
3	-0.29	0.07	2.54	0.04	2.30	10.24	1.04	7.72	23.92
4	-0.45	-0.44	0.11	-0.25	0.33	3.65	0.17	3.15	12.59
$n_1 = n_2 = 200$									
1	-0.20	-0.13	0.86	-0.13	0.80	6.38	0.44	5.36	24.67
2	-0.23	-0.21	-0.08	-0.20	-0.07	1.48	-0.09	1.34	8.52
3	-0.03	0.48	3.36	0.27	2.83	10.95	1.29	8.08	49.90
4	-0.07	0.01	0.82	-0.01	0.75	4.48	0.40	3.68	17.33

(though slightly less accurate) results as M1. On the other hand, the values of RPIs when M2 outperforms M1 usually range around $5\% \sim 25\%$ up to 49.90%, showing much more significant benefit. Therefore in the reward-risk trade off by using M2 instead of M1, the potential reward is more significant than risk.

4.3.2 Study 2: synthetic bathymetric surveys

In this study, we illustrate the proposed joint model of multiple surveys with a set of synthetic but realistic bathymetric surveys. The surveys and bathymetry data are synthesised using features from the real bathymetry databases. The main reason for which we employ synthetic data instead of real bathymetric surveys is that the "true" bathymetry surface can be assumed to be known for synthetic data so that it is easier to evaluate the model performance.

For illustration, we consider the region over $124.80°W \sim 125.10°W$ and $48.25°N \sim 48.55°N$ in front of the Strait of Juan de Fuca in the West coast of

Figure 4.8: Location of the study region near Strait of Juan de Fuca.

North America; see Figure 4.8. This region is of interest since the potential tsunami in the Cascadia area will possibly propagate through the strait and hit Victoria, the capital city of British Columbia, Canada. The bathymetry data are extracted from the 3 arc-second ($\sim$ 90 m) U.S. Coastal Relief Model (CRM) (NOAA NCEI, 2016).

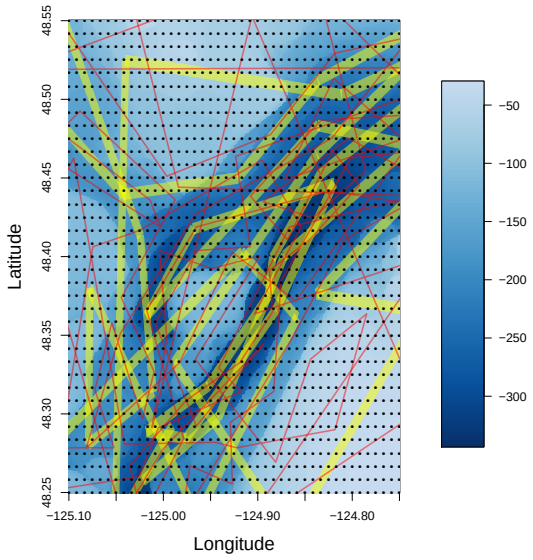


Figure 4.9: Synthetic surveys including 3 arc-second multibeam grids (yellow), low-resolution single beam or leadline surveys (red) and 30 arc-second sparse grid (black).

In order to mimic the realistic surveys, we consider three types of surveys as shown in Figure 4.9. The first type is the 3 arc-second multibeam echo sounder surveys. These surveys are usually of high resolution and accuracy because of the advanced multibeam technology. The multibeam sounders generate acoustic signals

through a wide angular lateral aperture transducer. Using the time for reflections of the lateral echoes of the sea bed that are received from multiples narrow beams, water depth can be extrapolated along a wide band. The width of the band should vary with the water depth. For simplicity, we assume a fixed band width here. The second type of surveys use leadlines or single beam echo sounders. Unlike multi-beam soundings, only water depth under the echo sounder base or leadline can be calculated so that these surveys usually contain much coarser data with probably larger errors. It is assumed that observations are taken every 6 arc-seconds along the track lines. The third type of surveys contain sparse data that come from various other sources such as digital soundings, digitized data from smooth sheets of hydrographic surveys and so on. This type of data are usually of high uncertainty and low resolution. It is represented by a 30 arc-second grid over the whole region. The measurement errors are simplified and assumed to be additive i.i.d Normal random variables with standard deviations of 2 m, 10 m and 20 m for the three types of surveys respectively. The synthetic data contain 44,615 measurement samples in total of which the data of the first, second and third type take up 84.22%, 12.22% and 3.56% respectively.

Then, we are able to fit the whole bathymetry surface for the study region with these data. Denoting the three sets of surveys as $(\mathbf{Y}_i, \mathbf{S}_i)$, $i = 1, 2, 3$ respectively and the underlying unknown spatial field as f_0, three models are considered.

M1: Letting $\mathbf{Y} = (\mathbf{Y}_1^T, \mathbf{Y}_2^T, \mathbf{Y}_3^T)^T$ and $\mathbf{S} = (\mathbf{S}_1^T, \mathbf{S}_2^T, \mathbf{S}_3^T)^T$, we assume $\mathbf{Y}|f_0 \sim \mathrm{N}(f_0(\mathbf{S}), \sigma_0^2)$ and $\mathbf{S}$ completely random.

M2: For $i = 1, 2, 3$, we assume that $\mathbf{Y}_i|f_0 \sim \mathrm{N}(f_0(\mathbf{S}_i), \sigma_i^2)$ and $\mathbf{S}_i$ completely random.

M3: For $i = 1, 2, 3$, we assume that $\mathbf{Y}_i|f_0 \sim \mathrm{N}(f_0(\mathbf{S}_i), \sigma_i^2)$, and for $i = 1, 2$, $\mathbf{S}_i \sim \mathrm{Poisson}(\exp\{a_i + b_i f_0(\mathbf{S}_i)\})$.

M1 is the simplest model that just combines all the data together. M2 and M3 are two variants of our proposed model (4.3) of which M2 assumes no preferential sampling while M3 assumes surveys 2 and 3 are possibly preferential. Taking the posterior means as predictions at the whole 3 arc-second grid over the study re-

gion, we compare the predictive performance of the three models by computing the normalised PRMSEs, that are 0.0353, 0.0284 and 0.0793 respectively. Therefore, by modelling the three surveys separately with M2 instead of M1, the normalised PRMSEs is reduced by around 20%. This is mainly due to the treatment of different measurement accuracy among the three types of surveys in M2. However, when preferential sampling feature is included as in M3, the prediction becomes much less accurate than the other two models. The posterior mean and standard deviation of the whole study region using the best model M2 are presented in Figure 4.10. The posterior mean surface is highly consistent with the true surface as shown in Figure 4.9 and captures the general features of this region. The posterior standard deviation surface provides us an assessment of the uncertainty in the predictions, which is currently not available in most of the public geo-database such as DEMs and DBMs. In general, the standard deviation is smaller over regions that are covered by bathymetric surveys than those without any surveys. The high standard deviation in regions where there are few or even no surveys suggests that one should be careful with the prediction as high uncertainty could be introduced into the following process, e.g. tsunami modelling, and proper uncertainty quantification is required.

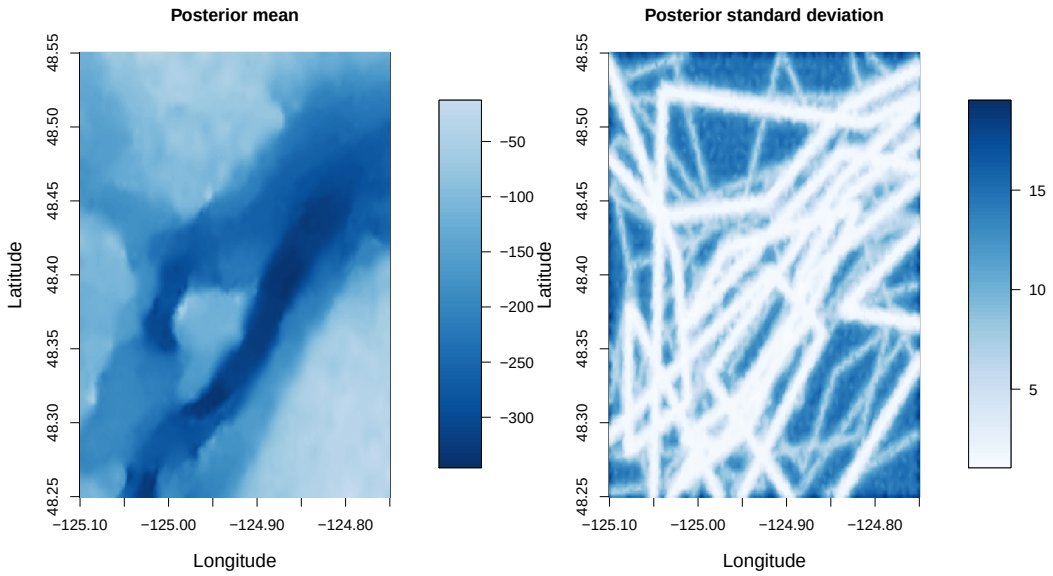


Figure 4.10: Posterior mean and standard deviation of the whole surface given three surveys using M2.

Though M3 yields the highest predictive error, the data seem to still suggest some preferential sampling feature. Figure 4.11 displays the posterior marginals of the parameters b_1 and b_2. Both parameters are significantly negative which implies that both survey 1 and 2 tend to be conducted in deep ocean. Figure 4.12 presents the predictive error over the whole study region using M3. The error could be ranging from -50 to 100 metres roughly. But M3 overestimates the field significantly over a large proportion of the study region, and most of these regions are not covered by any of the surveys. Since M3 tries to combine information from both measurements $\mathbf{Y}_i$ and the measured locations $\mathbf{S}_i$ for $i = 1, 2$, the lack of observations in those regions tends to imply relatively high values of f_0 because of the negative coefficients b_i.

This might explain why M3 makes mistakes in the prediction and overestimates the field in large area. This problem has warned us that the model for the preferential sampling using inhomogeneous Poisson process with intensity $\lambda(\mathbf{s}) = \exp\{a + bf_0(\mathbf{s})\}$ is probably too strong for the preferentiality in some real applications. Such explicit distributional assumption could help with the inference when the spatial point patterns actually or nearly follow the presumed distribution as shown in Section 4.3.1, but could be restrictive and misleading when the surveys are not perfectly sampled according to the distribution. For example, some bathymetry surveys could cover regions of 100 m to 2000 m deep with regular sampling locations regardless of the different water depth but with few or even no observations in regions of 50m or less. Overall, the surveys are obviously preferential as they avoid shallow water regions. But it is clear that the inhomogeneous Poisson assumption with intensity $\lambda(\mathbf{s}) = \exp\{a + bf_0(\mathbf{s})\}$ is not suitable to describe such preferentiality. On the other hand, the inhomogeneous Poisson model takes use of $f_0(\mathbf{s})$, which is however unknown and to be inferred, to define the intensity. If this assumption holds, it suggests that when conducting the surveys, the unknown latent field $f_0(\mathbf{s})$ is in fact known and used to generate the sampling locations. This counter-intuition also implies that the inhomogeneous Poisson model in this chapter could be too strong and unrealistic in some scenarios. It may be needed to impose some weak or

flexible prior on the relationship between the latent field values and point patterns to improve the inference with preferential sampling. We leave this for future work.

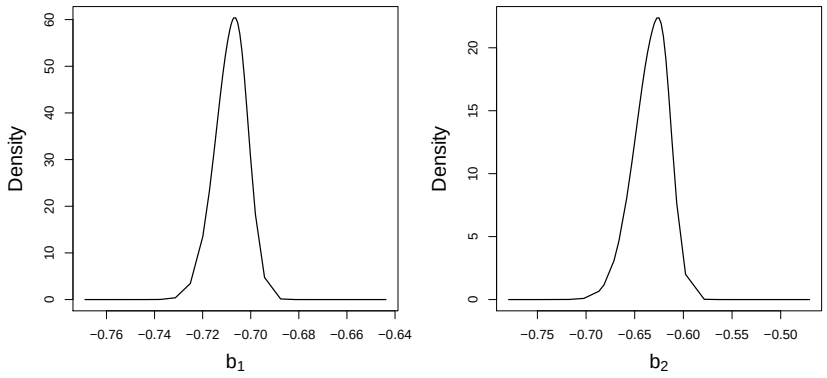


Figure 4.11: Posterior marginals of the parameters b_2 and b_3 using M3.

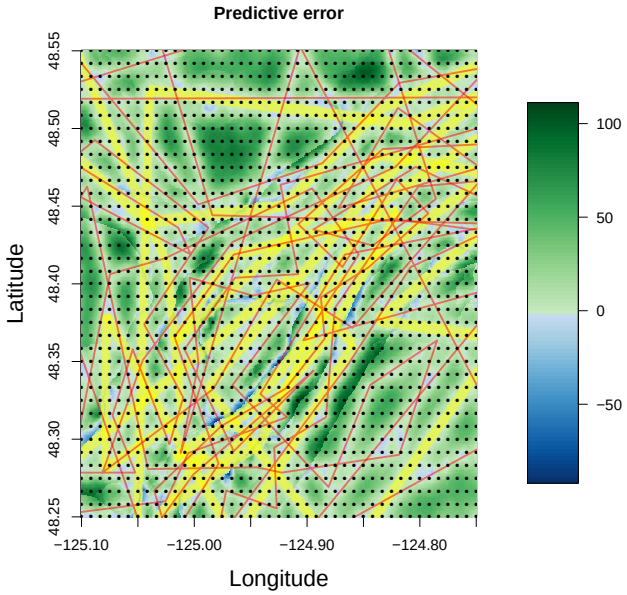


Figure 4.12: Predictive error of M3 with survey locations.

4.4 Discussion

In this chapter, we proposed a joint model to combine multiple spatial surveys based on the SPDE approach. The model aims to treat the surveys separately to account for their respective characteristics. We conducted a numerical simulation on Matérn

field under various situations to examine the proposed model in terms of parameter inference and spatial prediction. The model was demonstrated to be effective and yields superior results than the competitors, where the different characteristics of multiple surveys are neglected, in most cases. The joint model was then applied to a complicated and realistic bathymetry data set. It tackled the different bathymetric surveys effectively and improved the accuracy of the spatial prediction.

Despite the promising results, this work is just a proof-of-concept study for the issue of multiple spatial surveys. There are many aspects that can be explored further. Only spatial variations are considered here by assuming the latent field is fixed. However, the surveys are usually conducted over a long period and the underlying spatial process may evolve over time in some realistic applications. Such temporal evolution has been considered in the multivariate generalized linear model by Giorgi et al. (2015) for prevalence surveys. The seafloor is also changing due to many factors such as natural evolution and sediments. Hence, it might be helpful to include both space and time modelling in the joint model. The SPDE model has been demonstrated to be capable for spatial-temporal modelling; see Cameletti et al. (2013) for example. Moreover, the biased sampling problem arises often in many applications. Here, we follow Diggle et al. (2010) and assume a two-parameter representation using a inhomogeneous Poisson process to describe the potential preferential sampling feature. This assumption might be not realistic or sufficient to reveal the actual characteristics in some complicated applications, e.g. as shown in the application to bathymetry prediction in Study 2. Giorgi et al. (2015) noticed the problem where the bias is a function of the location itself rather than of the elevation. Therefore, more investigation need to be conducted to tackle the biased sampling issue.

Chapter 5

Conclusions and Future Work

5.1 Conclusion

In this thesis, two main topics have been discussed: uncertainty quantification of spatial fields and statistical emulation with high-dimensional input space. They are both motivated by tsunami modelling and the contributions are also illustrated with tsunami related applications.

The contributions to the first topic were described in Chapter 2 and 4. Specifically, we extended the SPDE approach by introducing bivariate splines in Chapter 2. This extension allows more flexible choices for the functional representation of the Gaussian field in latent Gaussian models. Instead of the conventional linear finite elements, we have implemented piecewise polynomial representations of various degrees using bivariate spline techniques. The improvement of our proposed method over the original method was highlighted through both theoretical exploration and intensive numerical studies. Various results have demonstrated that our approach using bivariate splines has enriched further the capability of the SPDE approach. In the numerical simulations, especially those with bathymetric data, we showed that the SPDE approach is effective in spatial prediction. Unlike many widely used mapping tools such as GMT, the SPDE approach tackles the uncertainties in a probabilistic way automatically and produces uncertainty estimates in the prediction. It was applied to the uncertainty quantification of bathymetry for tsunami modelling in Chapter 3.

The other issue in spatial modelling we discussed is the joint modelling of multiple surveys in Chapter 4. This issue arises often in bathymetric surveys where several surveys are usually conducted over a common region. It is advisable to combine the information from them. However, the surveys probably differ in many aspects such as survey trace, coverage and accuracy, and could be conducted preferentially due to the constraints in budget or safety. We proposed to tackle these different features in a joint latent Gaussian model, with separate layers for these surveys, based on the SPDE approach. The preferential sampling features could be included using log-Gaussian Cox processes. We have illustrated through numerical simulations with Matérn fields that the proposed joint model is effective and outperforms the alternative model where the differentiation among the surveys is neglected. This method was also applied into a bathymetry data set to improve the spatial prediction. These results have highlighted the need to take into account the respective characteristics of the surveys in spatial inference and prediction.

In Chapter 3, we discussed the second topic: statistical emulation for computer models with high-dimensional input space. This is motivated by the uncertainty propagation of the bathymetry through tsunami models where the bathymetry are represented as high-dimensional input, e.g. with finite elements or bivariate splines using the SPDE approach. Conventional statistical emulators are not able to deal with such problem. We proposed to merge the emulation with dimension reduction techniques. It was aimed to work on a low dimensional subspace that contains as much information for the input-output relationship as possible. A gradient-based kernel dimension reduction technique was employed for its wide capability and superior performance. Theoretical error bounds for such approximation were established. Numerical simulation on a PDE showed the advantages of the proposed emulation framework with the gKDR approach over a variety of other dimension reduction techniques in terms of effectiveness and efficiency. The key advantage of our method is its wide applicability in many scenarios involving various variable types without strong distributional assumptions and explicit calculation of gradients. We also applied the framework into the uncertainty quantification of tsunami

modelling with uncertain bathymetry. The simulation results have demonstrated the significant impact of the uncertainties in the bathymetry on tsunami waves. Therefore, it is plausible to include the uncertainties in the bathymetry into the tsunami uncertainty quantification and risk assessment process.

5.2 Future work

There are some potential applications of the work discussed in this thesis. For instance, the SPDE approach, together with the extension using bivariate splines and joint modelling of multiple spatial surveys, can be utilised in the statistical modelling of discrete bathymetry data in order to construct continuous bathymetry maps. These statistical methods account for the uncertainties in the observations and prediction so that they also provide uncertainty estimates of the bathymetry at a given location rather than a single value, that are not included in the current data products, e.g. DEM and GBD. Therefore, they can be applied into the bathymetry data production in order to improve the data quality and conduct the associated uncertainty quantification. In addition, the significant impact of the uncertainties in the bathymetry on tsunami waves has been demonstrated using GP emulation with dimension reduction. Thus it makes the tsunami hazard assessment more reliable and comprehensive to consider the spatial uncertainties in the bathymetry. This provides more information to the decision makers in order to evaluate the possible scenarios and make the right decision for various purposes including civil planning and emergency handling. Another practical implication could be within the Catastrophe models used by the (re)insurance industry to quantify the possible financial losses due to some hazards; see Appendix C for some practical discussion about the Catastrophe modelling of tsunami hazards and the associated financial impact of the uncertainties in the bathymetry.

There is also plenty of room for further methodological developments along the directions of this thesis: uncertainty quantification of various boundary conditions in complex computer models and the uncertainty propagation using statistical efficient methods.

We have shown that the SPDE approach using a Markov approximation is promising to quantify the uncertain boundary fields. We have also extended the SPDE by introducing bivariate splines of various orders instead of the linear finite elements and obtained more flexible representations of the data and efficient inference. However, some simulations have suggested that the higher order representations do not necessarily outperform their lower order counterparts. It needs to be investigated how to construct adaptive representations using bivariate splines with proper order and smoothness according to the specific data itself to strike a balance between model accuracy, computational operations and memory management, especially with the view of exploiting parallelization or GPU computing in high performance computing platforms.

Apart from the Markov approximation, there are some other directions to make use of Gaussian processes for complex applications to big data such as stochastic variational inference (Hoffman et al., 2013) and sparse GP with pseudo-inputs (Snelson and Ghahramani, 2005). Another possible approach for large scale data set is domain decomposition that makes inference for each sub-domain of moderate size and then these sub-domains are combined properly across the boundaries. These techniques and the other state-of-the-art developments could be also applied into the inference of bivariate splines for big data. Some boundary conditions are three-dimensional, e.g. the vertical atmospheric elevations of wind or temperature in climate models, the blood flow speed and pressure in biological models. The spherical and trivariate splines (Lai and Schumaker, 2007) could be introduced for these data. However, the extension from 2-D to 3-D may not be straightforward. For example, when Markov approximation is applied, the operational cost for factorising the precision matrix for a data set of n samples from Gaussian Markov field is $\mathcal{O}(n^2)$ for three dimensions (Rue et al., 2009) which could be still too expensive for big data. Therefore, more efficient methodologies need to be developed.

As discussed in Chapter 3, dimension reduction provides an intuitive way to tackle the high-dimensional emulation issue. However, the dimension reduction step itself usually requires sufficient samples to estimate the effective subspace ef-

ficiently. Proper practical criteria for the sample size required could be proposed based on the theoretical exploration of the error behaviour due to the reduced dimensions. When the computer models are expensive to run, how to make best use of the simulation runs within limited computational resource to estimate the effective subspace is also a key question to answer.

Simulation based approaches provide alternatives to uncertainty quantification with high-dimensional components as they directly analyse samples and do not require high-dimensional inference. While standard Monte Carlo simulation is too computationally expensive, multilevel Monte Carlo (MLMC) (Giles, 2008) approach can be applied. It requires to run the computer models at different resolutions from low to high. Due to the fact that most of the uncertainty can be captured by efficient low-resolution runs, MLMC dramatically reduces computational costs, with less high-resolution runs. The MLMC approach can be easily implemented and is effective to applications with high nonlinearity and dimensionality. It is also a promising method to handle the uncertainties in complex boundary conditions. MLMC may be integrated with dimension reduction to estimate the effective subspace based on a sequence of computer model runs from low to high resolutions instead of only a small number of high-resolution runs. When the dimension is reduced using dimension reduction techniques to a level that other emulation techniques still cannot afford, the MLMC approach can be applied for the uncertainty analysis. The relatively low-dimensional effective subspace may exclude the redundant information and make the inference using MLMC more effective and efficient.

Moreover, some computer models consist of multi-physics. For instance, a climate model can be a collection of various physical models for different atmospheric quantities. The uncertainties could be propagated across these models. For example, in a climate model, the uncertain sea surface temperature will influence the vertical atmospheric wind or pressure through a physical model, then the uncertainties in the wind or pressure will be propagated to the precipitation through another physical model, and so on. The uncertainty quantification across a system of multi-physics is also worthy of investigation.

Appendix A

Proofs for Chapter 2

A.1 Proof of Theorem 1

By plugging the bivariate spline representation of $x(\mathbf{u})$ in to the equality (2.4), we have

$$\left\{\langle\phi_t, \sum_{h=1}^{m}(\kappa^2-\Delta)^{\alpha/2}\tau\psi_h w_h\rangle, t=1,...,n_t\right\} \overset{\mathrm{d}}{=} \left\{\langle\phi_t, W\rangle, t=1,...,n_t\right\}, \quad \text{(A.1)}$$

for any appropriate set of test functions $\{\phi_t, t=1,...,n_t\}$.

When $\alpha = 1$:

By choosing a set of test functions to be $\phi_h = (\kappa^2-\Delta)^{1/2}\psi_h$, we have

$$\begin{aligned}&\left\{\langle(\kappa^2-\Delta)^{1/2}\psi_t, \sum_{s=1}^{m}(\kappa^2-\Delta)^{1/2}\tau\psi_s w_s\rangle, t=1,...,m\right\} \\ &\overset{\mathrm{d}}{=} \left\{\langle(\kappa^2-\Delta)\psi_t, W\rangle, t=1,...,m\right\}.\end{aligned} \quad \text{(A.2)}$$

Following Lemma 2 of Lindgren et al. (2011), the left hand side of (A.2) is

$$\left\{\sum_{s=1}^{m}\tau(\kappa^2\langle\psi_t,\psi_s\rangle + \langle\nabla\psi_t, \nabla\psi_s\rangle)w_s, t=1,...,m\right\},$$

when the Neumann boundary condition holds. The integral on the right hand side

is in fact Gaussian with mean zero and covariance matrix whose (t, s)-th element is

$$\begin{aligned}&\mathrm{Cov}(\langle(\kappa^2-\Delta)^{1/2}\psi_t, W\rangle, \langle(\kappa^2-\Delta)^{1/2}\psi_s, W\rangle)\\&= \langle(\kappa^2-\Delta)^{1/2}\psi_t, (\kappa^2-\Delta)^{1/2}\psi_s\rangle = \kappa^2\langle\psi_t, \psi_s\rangle + \langle\nabla\psi_t, \nabla\psi_s\rangle.\end{aligned}$$

Then we can write (A.2) in the matrix form as

$$\tau(\kappa^2\mathbf{M}+\mathbf{K})\mathbf{w} \sim N(\mathbf{0}, \kappa^2\mathbf{M}+\mathbf{K}),$$

where the (t, s)-th entry of the matrices $\mathbf{M}$, $\mathbf{K}$ are respectively $\mathbf{M}_{ts} = \langle\psi_t, \psi_s\rangle$, $\mathbf{K}_{ts} = \langle\nabla\psi_t, \nabla\psi_s\rangle$. $\mathbf{M}$ and $\mathbf{K}$ are usually named mass matrix and stiffness matrix respectively in bivariate spline literature. Therefore it is easy to show that the precision matrix of $\mathbf{w}$ is $\mathbf{Q} = \tau^2(\kappa^2\mathbf{M}+\mathbf{K})$.

When $\alpha = 2$:

We can choose the specific set of test functions to be $\phi_h = (\kappa^2-\Delta)\psi_h$ or $\phi_h = \psi_h$, leading to the least squares or Galerkin solutions respectively.

(1) When $\phi_h = (\kappa^2-\Delta)\psi_h$, we have

$$\begin{aligned}&\left\{\langle(\kappa^2-\Delta)\psi_t, \sum_{s=1}^{m}(\kappa^2-\Delta)\tau\psi_s w_s\rangle, t = 1, ..., m\right\}\\&\overset{\mathrm{d}}{=} \left\{\langle(\kappa^2-\Delta)\psi_t, W\rangle, t = 1, ..., m\right\}.\end{aligned} \tag{A.3}$$

The left hand side of (A.3) is

$$\left\{\sum_{s=1}^{m}\tau(\kappa^4\langle\psi_t, \psi_s\rangle + 2\kappa^2\langle\nabla\psi_t, \nabla\psi_s\rangle + \langle\Delta\psi_s, \Delta\psi_t\rangle)w_s, t = 1, ..., m\right\},$$

by applying the stochastic Green's first identity along with the Neumann boundary condition. The integral on the right hand side is Gaussian with mean zero and covariance matrix whose (t, s)-th element is

$$\mathrm{Cov}(\langle(\kappa^2-\Delta)\psi_t, W\rangle, \langle(\kappa^2-\Delta)\psi_s, W\rangle) = \kappa^4\langle\psi_t, \psi_s\rangle + 2\kappa^2\langle\nabla\psi_t, \nabla\psi_s\rangle + \langle\Delta\psi_s, \Delta\psi_t\rangle.$$

Then we can write (A.3) in the matrix form as

$$\tau(\kappa^4\mathbf{M} + 2\kappa^2\mathbf{K} + \mathbf{R})\mathbf{w} \sim N(\mathbf{0}, \kappa^4\mathbf{M} + 2\kappa^2\mathbf{K} + \mathbf{R}),$$

where the (t, s)-th entry of the matrix $\mathbf{R}$ is $\mathbf{R}_{ts} = \langle\Delta\psi_t, \Delta\psi_s\rangle$. The matrices $\mathbf{M}$ and $\mathbf{K}$ are defined above, and $\mathbf{R}$ is usually called roughness matrix. Then the precision matrix of $\mathbf{w}$ for the least squares solution can be easily shown to be $\mathbf{Q}^{LS} = \tau^2(\kappa^4\mathbf{M} + 2\kappa^2\mathbf{K} + \mathbf{R})$.

(2) When $\phi_h = \psi_h$, we have

$$\left\{\langle\psi_t, \sum_{s=1}^{m}(\kappa^2 - \Delta)\tau\psi_s w_s\rangle, t = 1, ..., m\right\} \overset{\mathrm{d}}{=} \left\{\langle(\kappa^2 - \Delta)\psi_t, W\rangle, t = 1, ..., m\right\}. \tag{A.4}$$

Following the same procedure as for the lease squares solution, we have the left hand side of (A.4) is in fact

$$\left\{\sum_{s=1}^{m}\tau(\kappa^2\langle\psi_t, \psi_s\rangle + \langle\nabla\psi_t, \nabla\psi_s\rangle)w_s, t = 1, ..., m\right\},$$

and the integral on the right hand side is Gaussian with mean zero and covariance matrix whose (t, s)-th element is

$$\mathrm{Cov}(\langle\psi_t, W\rangle, \langle\psi_s, W\rangle) = \langle\psi_t, \psi_s\rangle.$$

Then (A.4) can be re-written as

$$\tau(\kappa^2\mathbf{M} + \mathbf{K})\mathbf{w} \sim N(\mathbf{0}, \mathbf{M}).$$

Then the precision matrix of $\mathbf{w}$ for the Galerkin solution is $\mathbf{Q}^G = \tau^2(\kappa^2\mathbf{M} + \mathbf{K})\mathbf{M}^{-1}(\kappa^2\mathbf{M} + \mathbf{K}) = \tau^2(\kappa^4\mathbf{M} + 2\kappa^2\mathbf{K} + \mathbf{K}\mathbf{M}^{-1}\mathbf{K})$.

When $\alpha \geq 3$:

Following the recursive algorithm, we can find the solution to the SPDE $\tau(\kappa^2 - \Delta)^{\alpha/2}x(\mathbf{u}) = W(\mathbf{u})$ by solving the innovative SPDE $(\kappa^2 - \Delta)x(\mathbf{u}) = \tilde{x}(\mathbf{u})$, where

$\tilde{x}(\mathbf{u})$ is the solution to the SPDE $\tau(\kappa^2 - \Delta)^{(\alpha-2)/2}\tilde{x}(\mathbf{u}) = W(\mathbf{u})$. Then by choosing the test functions $\phi_h = \psi_h$, $h = 1, ..., m$ and following the same procedure for the Galerkin solution when $\alpha = 2$, we have

$$\mathbf{Q}_\alpha = (\kappa^2\mathbf{M} + \mathbf{K})\mathbf{M}^{-1}\mathbf{Q}_{\alpha-2}\mathbf{M}^{-1}(\kappa^2\mathbf{M} + \mathbf{K}),$$

which can be expanded to the expression in the theorem.

Then we show the calculations of the matrix components $\mathbf{M}$, $\mathbf{K}$ and $\mathbf{R}$. Following Lemma 1 and $\nabla p = \sum_{i+j+k=d} c_{ijk}\nabla B^d_{ijk}$, $\Delta p = \sum_{i+j+k=d} c_{ijk}\Delta B^d_{ijk}$ for any $p = \sum_{i+j+k=d} c_{ijk}B^d_{ijk}$, we have the contribution of triangle T to the (t, s)-th entry of $\mathbf{M}$, $\mathbf{K}$ and $\mathbf{R}$ for $t, s = 1, ..., m$ are

$$\mathbf{M}_{ts}|_T = \langle\psi_t, \psi_s\rangle_T = \mathbf{c}'_t|_T\mathbf{M}_T\mathbf{c}_s|_T,$$

$$\mathbf{K}_{ts}|_T = \langle\nabla\psi_t, \nabla\psi_s\rangle_T = \mathbf{c}'_t|_T\mathbf{K}_T\mathbf{c}_s|_T,$$

$$\mathbf{R}_{ts}|_T = \langle\nabla\psi_t, \nabla\psi_s\rangle_T = \mathbf{c}'_t|_T\mathbf{R}_T\mathbf{c}_s|_T,$$

where $\mathbf{M}_T$, $\mathbf{K}_T$ and $\mathbf{R}_T$ are defined in Theorem 1, and $\mathbf{c}_h|_T$ is the column vector of B-coefficients of ψ_h associated with triangle T, $h = 1, ..., m$. Then it is followed that $\mathbf{M}_{ts} = \sum_T \mathbf{M}_{ts}|_T = \mathbf{c}'_t\mathbf{M_0}\mathbf{c}_s$, $\mathbf{K}_{ts} = \sum_T \mathbf{K}_{ts}|_T = \mathbf{c}'_t\mathbf{K_0}\mathbf{c}_s$ and $\mathbf{R}_{ts} = \sum_T \mathbf{M}_{ts}|_T = \mathbf{c}'_t\mathbf{R_0}\mathbf{c}_s$, where $\mathbf{M_0} = \mathrm{diag}(\mathbf{M}_T, T \in \mathbf{\Delta})$, $\mathbf{K_0} = \mathrm{diag}(\mathbf{K}_T, T \in \mathbf{\Delta})$ and $\mathbf{R_0} = \mathrm{diag}(\mathbf{R}_T, T \in \mathbf{\Delta})$. Therefore we have the following simple matrix representation that

$$\mathbf{M} = \mathbf{C}'\mathbf{M_0}\mathbf{C}, \quad \mathbf{K} = \mathbf{C}'\mathbf{K_0}\mathbf{C}, \quad \mathbf{R} = \mathbf{C}'\mathbf{R_0}\mathbf{C}.$$

A.2 Proof of Proposition 1

Let $f_{\mathbf{\Delta}}(\mathbf{s})$ be the H^1-orthogonal projection of $f \in H^1 \cap W_2^{m+1}(\Omega)$ onto the bivariate spline space $S^0_d(\mathbf{\Delta})$, it follows that

$$\int_\Omega f(\mathbf{s})Lx_{\mathbf{\Delta}}(\mathbf{s})\,\mathrm{d}\,\mathbf{s} = \int_\Omega (f(\mathbf{s}) - f_{\mathbf{\Delta}}(\mathbf{s}))Lx_{\mathbf{\Delta}}(\mathbf{s})\,\mathrm{d}\,\mathbf{s} + \int_\Omega f_{\mathbf{\Delta}}(\mathbf{s})Lx_{\mathbf{\Delta}}(\mathbf{s})\,\mathrm{d}\,\mathbf{s}$$

$$= \int_\Omega f_{\mathbf{\Delta}}(\mathbf{s}) L x_{\mathbf{\Delta}}(\mathbf{s}) \,\mathrm{d}\,\mathbf{s}$$
$$= \int_\Omega f_{\mathbf{\Delta}}(\mathbf{s}) \,\mathrm{d}\, W(\mathbf{s}),$$

where the second equality follows from the orthogonality of $f(\mathbf{s}) - f_{\mathbf{\Delta}}(\mathbf{s})$ to $S_d^0(\mathbf{\Delta})$ with respect to H^1 inner product. Then we have

$$\int_\Omega f(\mathbf{s}) L(x(\mathbf{s}) - x_{\mathbf{\Delta}}(\mathbf{s})) \,\mathrm{d}\,\mathbf{s} = \int_\Omega (f(\mathbf{s}) - f_{\mathbf{\Delta}}(\mathbf{s})) \,\mathrm{d}\, W(\mathbf{s}).$$

Hence it follows from the white noise integrals that

$$\mathrm{E}\left(\int_\Omega f(\mathbf{s}) L(x(\mathbf{s}) - x_{\mathbf{\Delta}}(\mathbf{s})) \,\mathrm{d}\,\mathbf{s}\right)^2 = \mathrm{E}\left(\int_\Omega (f(\mathbf{s}) - f_{\mathbf{\Delta}}(\mathbf{s})) \,\mathrm{d}\, W(\mathbf{s})\right)^2$$
$$= \int_\Omega (f(\mathbf{s}) - f_{\mathbf{\Delta}}(\mathbf{s}))^2 \,\mathrm{d}\,\mathbf{s}.$$

Then it follows from standard results in bivariate splines literatures, for example Th. 5.19 in Lai and Schumaker (2007) that under some suitable assumptions on the triangulation, we have for $1 \le m \le d$,

$$\|f - f_{\mathbf{\Delta}}\|_{2,\Omega} \le K |\mathbf{\Delta}|^{m+1} |f|_{m+1,2,\Omega}.$$

A.3 Proof of Proposition 2

This proof was completed by M.-J. Lai. First of all, it is easy to see that

$$w_f' \mathbf{M} w_g = \langle f_{\mathbf{\Delta}}, g_{\mathbf{\Delta}} \rangle_{\mathbf{\Delta}} = \sum_{T \in \mathbf{\Delta}} \int_T f_{\mathbf{\Delta}} g_{\mathbf{\Delta}} \,\mathrm{d}\, x \,\mathrm{d}\, y \tag{A.5}$$

since $f_{\mathbf{\Delta}}, g_{\mathbf{\Delta}} \in S_d^0(\mathbf{\Delta})$. Next we can see

$$\mathbf{w}_f' \tilde{\mathbf{M}} \mathbf{w}_g = \sum_{T \in \mathbf{\Delta}} \sum_{\xi \in \mathcal{D}_{d,T}} c_\xi(f_{\mathbf{\Delta}}) \sum_\eta \int_T \phi_\xi \phi_\eta \,\mathrm{d}\, x \,\mathrm{d}\, y c_\xi(g_{\mathbf{\Delta}})$$
$$= \sum_{T \in \mathbf{\Delta}} \sum_{\xi \in \mathcal{D}_{d,T}} \frac{A_T}{\binom{d+2}{2}} c_\xi(f_{\mathbf{\Delta}}) c_\xi(g_{\mathbf{\Delta}}),$$

where $\mathcal{D}_{d,T} = \{(i\mathbf{v}_1 + j\mathbf{v}_2 + k\mathbf{v}_3)/d, i+j+k=d\}$ is the set of associated domain points of triangle $T = \langle \mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 \rangle$, A_T is the area of triangle T and $c_\xi(s)$ is the B-coefficient of s. When $f_{\boldsymbol{\Delta}} = C$ is a constant C, it is easy to see that

$$\sum_{\xi \in \mathcal{D}_{d,T}} \frac{A_T}{\binom{d+2}{2}} c_\xi(f_{\boldsymbol{\Delta}}) c_\xi(g_{\boldsymbol{\Delta}}) = \int_T f_{\boldsymbol{\Delta}} g_{\boldsymbol{\Delta}} \,\mathrm{d}\, x \,\mathrm{d}\, y$$

and hence, we have

$$\mathbf{w}_f' \tilde{\mathbf{M}} \mathbf{w}_g = \sum_{T \in \boldsymbol{\Delta}} \int_T f_{\boldsymbol{\Delta}} g_{\boldsymbol{\Delta}} \,\mathrm{d}\, x \,\mathrm{d}\, y$$

which is $w_f' \mathbf{M} w_g$ by (A.5). Similar when $g_{\boldsymbol{\Delta}}$ is a piecewise constant. Also, when $d = 1$, this result follows from Lemma 1 in Chen and Thomée (1985). We now prove it for general $d \geq 1$.

We first note that

$$\begin{aligned} c_\xi(f_{\boldsymbol{\Delta}}) c_\xi(g_{\boldsymbol{\Delta}}) =& (c_\xi(f_{\boldsymbol{\Delta}}) - f_{\boldsymbol{\Delta}}(\xi)) c_\xi(g_{\boldsymbol{\Delta}}) + f_{\boldsymbol{\Delta}}(\xi)(c_\xi(g_{\boldsymbol{\Delta}}) - g_{\boldsymbol{\Delta}}(\xi)) \\ &+ f_{\boldsymbol{\Delta}}(\xi) g_{\boldsymbol{\Delta}}(\xi). \end{aligned}$$

Then we claim that

$$\sum_{\xi \in \mathcal{D}_{d,T}} \frac{A_T}{\binom{d+2}{2}} f_{\boldsymbol{\Delta}}(\xi) g_{\boldsymbol{\Delta}}(\xi) \text{ approximates } \int_T f_{\boldsymbol{\Delta}}(x,y) g_{\boldsymbol{\Delta}}(x,y) \,\mathrm{d}\, x \,\mathrm{d}\, y.$$

Indeed, let us recall the Bernstein-Bézier approximation of arbitrary continuous function F on T. That is, using Th. 2.45 in Lai and Schumaker (2007), we have

$$\|F - B_d(F)\|_{T,\infty} \leq \frac{|T|^2}{d} |F|_{2,T} \tag{A.6}$$

where $B_d(F) = \sum_{\xi \in \mathcal{D}_{d,T}} F(\xi) B_\xi$ and B_ξ are the Bernestein-Bézier polynomials of degree d. Letting $F(x,y) = f_{\boldsymbol{\Delta}}(x,y) g_{\boldsymbol{\Delta}}(x,y)$, we have

$$\begin{aligned} &\Big| \int_T F(x,y) \,\mathrm{d}\, x \,\mathrm{d}\, y - \int_T B_d(F) \,\mathrm{d}\, x \,\mathrm{d}\, y \Big| \\ =& \Big| \int_T f_{\boldsymbol{\Delta}}(x,y) g_{\boldsymbol{\Delta}}(x,y) \,\mathrm{d}\, x \,\mathrm{d}\, y - \sum_{\xi \in \mathcal{D}_{d,T}} f_{\boldsymbol{\Delta}}(\xi) g_{\boldsymbol{\Delta}}(\xi) \frac{A_T}{\binom{d+2}{2}} \Big| \end{aligned}$$

$$
\begin{aligned}
\leq & \frac{|T|^2}{d} \int_T |f_{\boldsymbol{\Delta}} g_{\boldsymbol{\Delta}}|_{2,T} \,\mathrm{d}\, x \,\mathrm{d}\, y \\
\leq & K \frac{|T|^2}{d} |f_{\boldsymbol{\Delta}} g_{\boldsymbol{\Delta}}|_{2,1,T} \\
\leq & K \frac{|T|^2}{d} |f_{\boldsymbol{\Delta}}|_{2,2,T} |g_{\boldsymbol{\Delta}}|_{2,2,T},
\end{aligned}
$$

where we have used the fact that $f_{\boldsymbol{\Delta}} g_{\boldsymbol{\Delta}}$ is a polynomial of degree $2d$ in the second inequality and the Cauchy-Schwarz inequality in the last inequality. This finishes the proof of the claim.

Next we consider

$$
I_1(T) := \sum_{\xi \in \mathcal{D}_{d,T}} \frac{A_T}{\binom{d+2}{2}} (f_{\boldsymbol{\Delta}}(\xi) - c_\xi(f_{\boldsymbol{\Delta}})) c_\xi(g_{\boldsymbol{\Delta}}).
$$

We have

$$
|I_1(T)| = A_T \|\{f_{\boldsymbol{\Delta}}(\xi) - c_\xi(f_{\boldsymbol{\Delta}})\}_{\xi \in \mathcal{D}_{d,T}}\|_\infty \|\{c_\xi(g_{\boldsymbol{\Delta}})\}_{\xi \in \mathcal{D}_{d,T}}\|_\infty
$$

and hence, by Th. 2.6 in Lai and Schumaker (2007),

$$
|I_1(T)| \leq A_T K^2 \|B_d(f_{\boldsymbol{\Delta}}) - f_{\boldsymbol{\Delta}}\|_T |g_{\boldsymbol{\Delta}}|_T,
$$

where K is a positive constant. We use the property of Bernstein-Bézier approximation again, i.e. the estimate in (A.6) to have

$$
\begin{aligned}
|I_1(T)| \leq & K^2 A_T \frac{|T|^2}{d} |f_{\boldsymbol{\Delta}}|_{2,T} \|g_{\boldsymbol{\Delta}}\|_{\infty,\Omega} \\
\leq & K^2 \|f_{\boldsymbol{\Delta}}\|_{2,1,T} \|g_{\boldsymbol{\Delta}}\|_{\infty,\Omega}.
\end{aligned}
$$

Therefore we have

$$
\sum_{T \in \boldsymbol{\Delta}} |I_1(T)| \leq K^2 \frac{|T|^2}{d} \|g_{\boldsymbol{\Delta}}\|_{\infty,\Omega} \|f_{\boldsymbol{\Delta}}\|_{2,1,\Omega}.
$$

Similarly, we can discuss

$$I_2(T) := \sum_{\xi \in \mathcal{D}_{d,T}} \frac{A_T}{\binom{d+2}{2}} f_{\mathbf{\Delta}}(\xi)(c_\xi(g_{\mathbf{\Delta}}) - g_{\mathbf{\Delta}}(\xi))$$

to have a similar estimate as $I_1(T)$. Putting these three estimates above, we have obtained

$$|\epsilon_{\mathbf{\Delta}}(f_{\mathbf{\Delta}}, g_{\mathbf{\Delta}})| \leq K|\mathbf{\Delta}|^2(\|f_{\mathbf{\Delta}}\|_{2,2,\Omega}\|g_{\mathbf{\Delta}}\|_{2,2,\Omega} + \|f_{\mathbf{\Delta}}\|_{2,1,\Omega}\|g_{\mathbf{\Delta}}\|_{\infty,\Omega} + \|f_{\mathbf{\Delta}}\|_{\infty,\Omega}\|g_{\mathbf{\Delta}}\|_{2,1,\Omega}),$$

where K is a positive constant, $|\mathbf{\Delta}|$ is the length of the longest edge in the triangulation $\mathbf{\Delta}$. These complete the proof.

Appendix B

Contribution to the UCLB project on tsunami risk assessment in Cascadia

This is a joint work with Serge Guillas, Simon Day and Andria Sarri as part of the UCL-Business funded proof-of-concept project. The project is to quantify tsunami risk in Cascadia. The region is located at the Pacific Northwest of North America, as shown in Figure B.1, where located some well known cities like Victoria, Seattle and Vancouver. The project together with a follow-up project lead to an implementation of the UCL Cascadia tsunami hazard model which is discussed in the next appendix. I was responsible for the data acquisition of integrated bathymetry and topography, triangular mesh construction and initial tsunami runs using VOLNA on Emerald (one of the largest GPU clusters in Europe).

Based on the characteristics of the fault system in Cascadia and the relevant geophysical expertise, a complex seabed deformation mechanism is proposed over an irregular shaped realistic fault zone. The displacement motion, which can be represented with a few parameters, starts from the North and moves to the South with different maximum uplifts and spread speeds. Figure B.2 shows the seabed uplift after 182 seconds from the beginning for one of the seabed displacement designs. The computational domain is extended towards the West to avoid the refection effect.

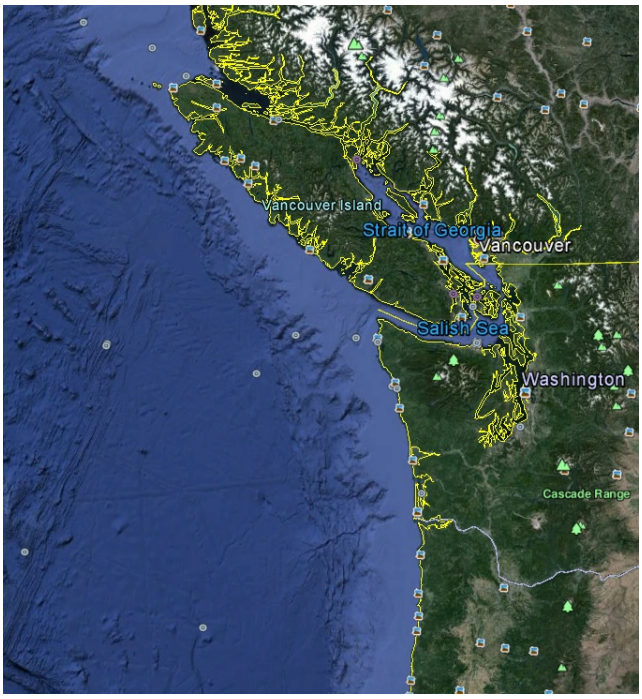


Figure B.1: Snapshot of the Cascadia area from Google Earth.

B.1 Merge DEMs of bathymetry and topography for Cascadia region

To simulate high-resolution tsunami events for this area, integrated bathymetry and topography data are required. There are many data products with different resolutions available in NOAA. Since we are more interested in the tsunami waves that move towards the coast and the inundation on land, more information about the bathymetry and topography in the coastal area is needed. But in the deep ocean in the West, high-resolution data are not necessary because of the little influence on the tsunami propagation towards the land. In particular, the following five DEMs are merged together in our simulation:

DEM 1. ETOPO1 Global Relief Model 1 arc-minute
http://www.ngdc.noaa.gov/mgg/global/

DEM 2. Strait of Juan de Fuca, WA 30 arc-second MHW DEM
http://www.ngdc.noaa.gov/dem/squareCellGrid/download/541

DEM 3. Strait of Juan de Fuca, WA 5 arc-second MHW DEM

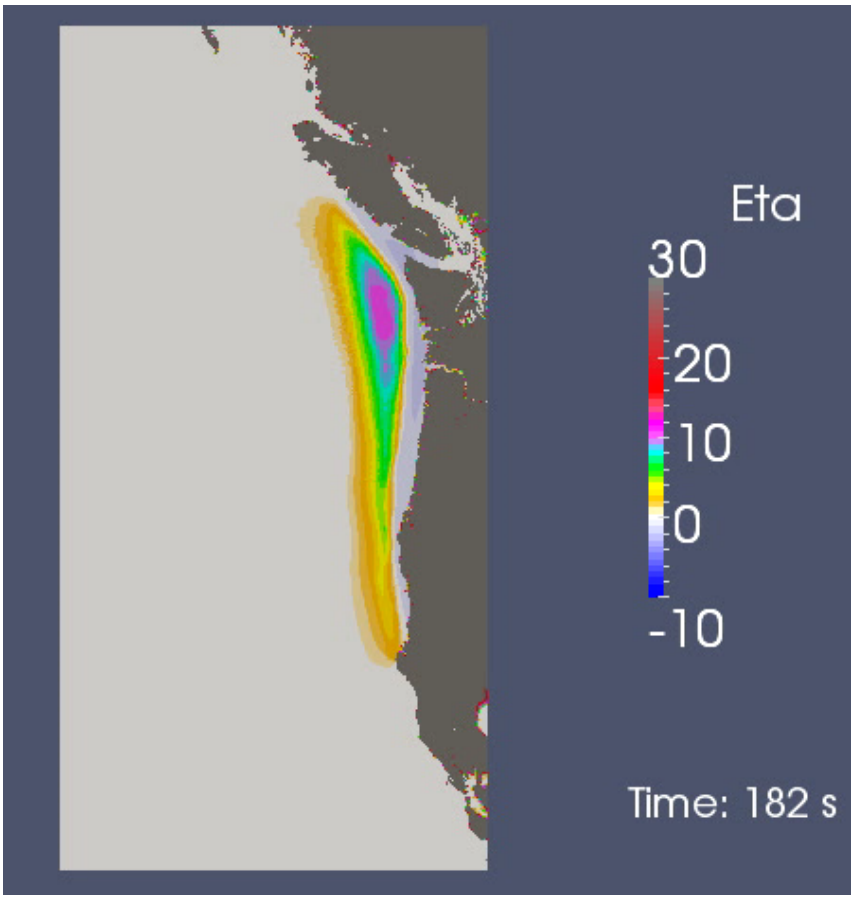


Figure B.2: An example of the seabed displacement in Cascadia at 182 seconds after the tsunami generation.

http://www.ngdc.noaa.gov/dem/squareCellGrid/download/655

DEM 4. US Coastal Relief Model - Northwest Pacific 3 arc-sec

http://www.ngdc.noaa.gov/mgg/coastal/grddas08/grddas08.htm

DEM 5. US Coastal Relief Model - Central Pacific 3 arc-sec

http://www.ngdc.noaa.gov/mgg/coastal/grddas07/grddas07.htm

Some DEMs are recorded in the Geographical coordinate system (latitude and longitude). They are converted into a common coordinate system, Washington State Plane South, that is a simple Cartesian coordinate system in meters, as required by the VOLNA code. The spatial coverage of these DEMs are shown in Figure B.3. The DEM 1 (etopo1) has the largest spatial coverage and the lowest resolution ($\sim$ 1800 m). The resolutions of DEM 2 (sjdf30) and DEM 3 (sjdf5) are about 900 m

and 150 m respectively. The other two DEMs, denoted by CRM nwp and CRM cp, have the highest resolution which is roughly 90 m. We can see that most of the coast area can be covered by DEMs with relatively high resolution.

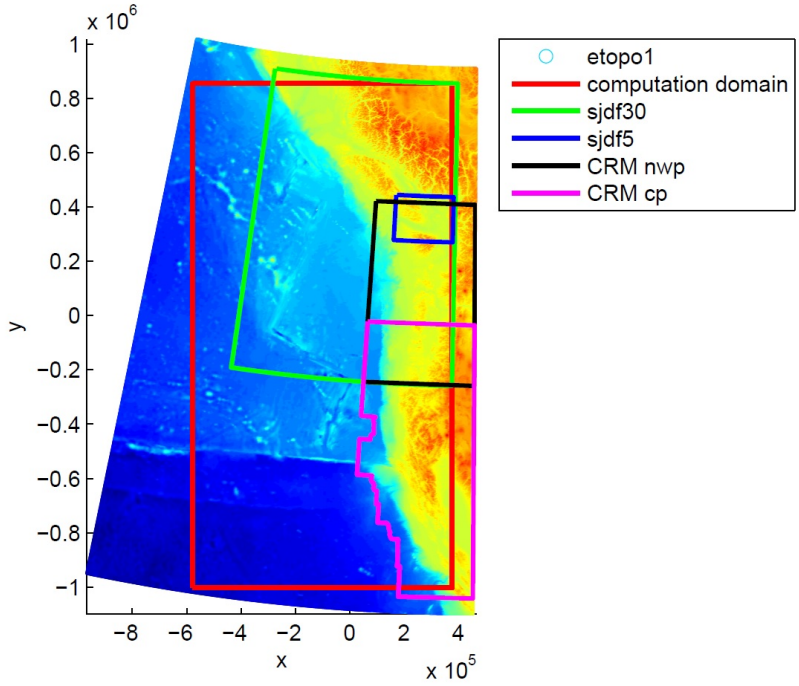


Figure B.3: DEMs merged for Cascadia.

To merge the five DEMs into a new data set, a simple rule is applied: where higher resolution data are available, we use higher resolution data instead of lower resolution data. The procedure can be described as:

for each DEM i, $i = 1, 2, 3, 4, 5$:
  for each data point (x, y, z) in the DEM i:
    if (x, y) is covered by DEMs with resolution higher than DEM i:
      go to next data point in DEM i;
    else:
      stack the data point (x, y, z) into the new data set.

Iterating over all the data points in the five DEMs, we get a merged new data set which has 77,113,088 data points in total over the computation domain.

B.2 Unstructured triangular mesh construction

The VOLNA code employs the finite volume method to solve the governing equations of a tsunami numerically over an unstructured triangular mesh. An appropriate mesh needs to be constructed for high-resolution tsunami simulations to strike a balance between the resolution and the computational cost.

The usual wind-generated waves on the beach often have a period (the time between two successional waves) of only a few seconds and a wavelength of about 100 to 200 metres. But a tsunami can have a period from 10 minutes up to 2 hours and a wavelength of more than 500 kilometres in the deep water. Because of the long wavelength, tsunami waves are characterised as shallow-water waves. For shallow-water waves, the speed can be roughly derived as $v = \sqrt{gh}$, where g is the acceleration of gravity and h is the water depth. Therefore, a tsunami usually travels at high speed in the deep water but gets slower as moving into the shallow water. Then the period decreases and the wavelength is reduced to $100 \sim 200$ metres roughly at the beach. Therefore, we need dense triangles in the shallow water to capture the subtle movement of tsunami waves and coarse triangles in the deep ocean to reduce the computing cost. In addition, more triangles are placed in regions where the gradients of the water depth are relatively large because sharp changes in the water depth might influence the wave propagation significantly. The specific rule for the specification of the mesh size is described as below.

- For area covered by high resolution DEMs (sjdf5, CRM-nwp, CRM-cp):
  - Below the sea level: if water depth for some area is greater than 250 m, set mesh size to be 1000 m; if it is smaller than 10 m, set mesh size to be 200 m and use linear interpolation to calculate the mesh size for area with water depth between 250 m and 10 m.
  - Above the sea level: if topography for some area is greater than 50 m, set mesh size to be 1000 m; if it is smaller than 10 m, set mesh size to be 200 m and use linear interpolation to calculate the mesh size for area with topography between 50 m and 10 m.

- For area covered by low resolution DEMs (sjdf30, ETOPO1):
  - Below the sea level: if water depth for some area is greater than 3000 m, set mesh size to be 15000 m; if it is smaller than 150 m or gradient is greater than 90% upper percentile of all gradients (about 0.77 here), set mesh size to be 3000 m and use linear interpolation to calculate the mesh size for area with water depth between 3000 m and 150 m and gradient less than 0.77.
  - Above the sea level: if topography for some area is greater than 100 m, set mesh size to be 15000 m; if it is smaller than 50 m, set mesh size to be 3000 m and use linear interpolation to calculate the mesh size for area with topography between 100 m and 50 m.
- For the extended area in the West, set mesh size to be 100000 m to save computational cost.

Based on such a rule of generating the triangulation, a mesh is constructed as shown in Figure B.4 depending on the water depths, gradients and the densities of observations. The mesh has 2,392,352 triangles in total. Then, the data points in the merged new data set as described in the previous section are mapped onto the barycentre of each triangle in the mesh to represent the initial bathymetry and topography for the whole computation domain in VOLNA.

B.3 Initial study of tsunami risk over Grays Harbor

To prepare for the tsunami risk assessment over the whole Cascadia region, an initial small-scale study over the Grays Harbor is conducted. The tsunami simulation values are output at the equally spaced gauge sites on a grid of 200 m $\times$ 200 m over the area as shown in Figure B.5.

Figure B.6 shows the inundation maps of the maximum flooding depth over the Grays Harbor area generated from the tsunami simulations using three different seabed displacement designs. The three events can be summarised by the size of the tsunami: large, middle and small. The possible inundation around the coast is

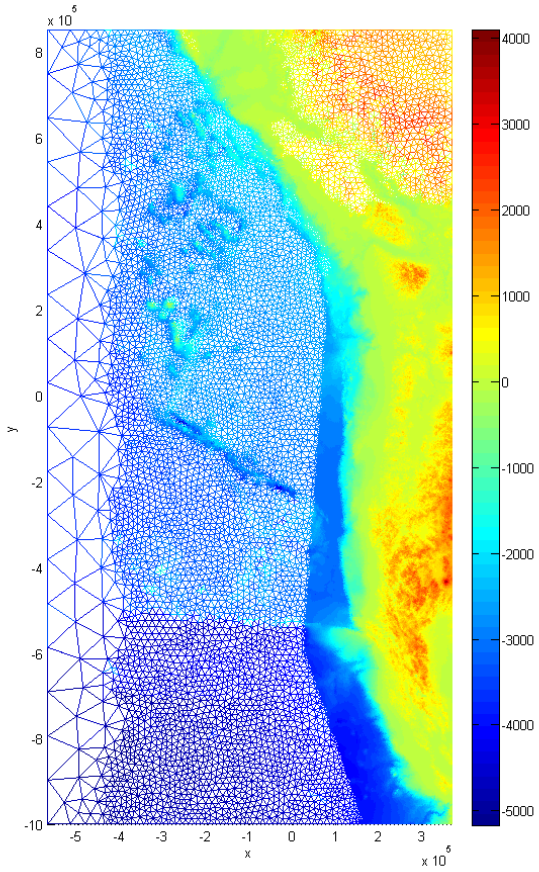
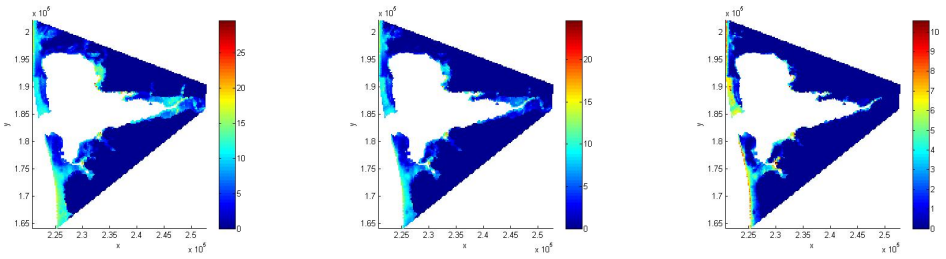


Figure B.4: Mesh for the Cascadia area (unit: metres).

Figure B.5: Location of the Grays Harbor in Google Map.

clearly significant in all of the three events. Furthermore, the inundation information can be written into *.kml* files so that we are able to visualise the inundation in Google Earth. For example, as shown in Figure B.7, a house in the middle could be inundated by the tsunami waves.



(a) event 1 (b) event 2 (c) event 3

Figure B.6: Inundation maps over the Grays Harbor of three tsunami events.

Figure B.7: Inundation visualised in Google Street View.

Appendix C

Combining the tsunami hazard model with the Oasis LMF Catastrophe modelling platform

This appendix is devoted to the commercial evaluation of some research findings with a primary focus on the improved tsunami hazard model that has promising commercial values in the Catastrophe modelling market. In Section C.1 and C.2, we give a brief introduction to the Catastrophe models and the market. Then we focus on the Oasis platform in Section C.3, which is a newly developed open marketplace for Catastrophe modelling. In Section C.4, we conduct several numerical studies to evaluate the calculation performance of the Oasis platform using the UCL Cascadia tsunami hazard model. Then we highlight the potential impact of the uncertainties in the bathymetry on tsunami hazards and hence the financial losses using the Oasis platform in Section C.5. This work is funded by the UCL Advances Enterprise Scholarship.

C.1 Catastrophe modelling market

Catastrophe modelling, or Cat modelling, is the process of using computer-assisted calculations to estimate or predict the losses due to a catastrophic event such as a hurricane, earthquake or tsunami. It is widely used in many applications. Insurance companies use Cat models for risk assessment of a portfolio of exposures, which

helps with the underwriting strategy, purchase of reinsurance or calculation of the premium to charge their policyholders. The financial strength and status of insurance companies that take catastrophic risk could also be assessed with Cat models by external rating agencies such as A. M. Best and Standard & Poor's. Reinsurers use Cat models in the pricing and structuring of their reinsurance polices. In the EU, because of the Solvency II regulations insurance companies need also derive the required regulatory capital using Cat models.

The market of Cat models was recently estimated to be worth around £400 million per annum with the "big three" leading players being AIR, RMS and EQECAT. Cat models have historically been proprietary or "black box" solutions whose operations and underlying models were not visible to insurance industry subscribers and where development was largely undertaken internally within the Cat modelling company. However, the new EU Solvency II regulations that is introduced since January 2013 require insurance companies to display a quantitative understanding of the risks to which they are exposed by their sales of insurance products, including an understanding of the uncertainties that propagate through the models and into the outputs. As a result, an industry movement towards the development of more transparent or "open" Cat models has gained traction, within which the Oasis consortium is attracting most interest. Additionally, the insurance industry is currently working with the Association for Cooperative Operations Research and Development (ACORD) to develop an industry standard for collecting and sharing exposure data, which are currently closed and proprietary. All of these changes result in an urgent need for new and advanced Cat models and hence potentially speedy growth of the Cat modelling market.

Among the various catastrophic hazards that are covered by Cat modelling, tsunami hazards haven't been addressed extensively until the recent tsunami disasters, most notably the Tōhoku 2011 earthquake-generated tsunami. These extreme tsunami events have highlighted the large potential losses to which the insurance industry is exposed through its cover of coastal portfolios in important tsunami-prone regions such as Japan and Cascadia (NW United States of America and Pacific

Canada). For example, EQECAT initially put the loss that the (re)insurance market would be liable for at between $12 billion and $25 billion, but within two months of the 2011 Tōhoku event they had increased their estimate to between $22 billion and $39 billion. The Tōhoku earthquake and tsunami have thus brought home to the insurance industry the importance of accurate and reliable tsunami risk modelling, coupled with accurate and reliable seismic shaking risk models to estimate catastrophe exposures to future similar disasters, and therefore the consequences for the entire insurance industry. These developments have resulted in a significant period of change in the insurance marketplace and provided significant translation and commercialisation opportunities for organisations such as universities who are developing novel or improved hazard modelling capabilities. The Oasis platform bridges such industry-academia connection through its unique and open platform. Both model providers and end users could benefit from the knowledge transfer in an easy and secure way.

C.2 Overview of Cat models

We may describe a Cat model with roughly four essential modules, though they may present in different forms in specific Cat models. They are exposure module, hazard module, vulnerability module and financial module.

Exposure module describes the exposures that are assessed against each hazard event in a Cat model. It consists of various features about each exposure such as the location, insured value, and the associated insurance policy terms. There may also be some other characteristics about the exposure buildings including construction type and year built to further assess the vulnerability of specific exposure against the hazard risk.

The hazard module contains the description of the potential hazards such as the frequency, likelihood and intensity. Examples include the wind speeds and pressure for hurricane, the epicentre and magnitude for earthquake. These characteristics are usually defined at each location over the region that is covered by a Cat model. More specifically, they are provided as a catalogue of events that are simulated and

produced with physical and statistical models. For each event, the likelihood or frequency it may strike is assessed, as well as the the potential hazard intensity at each location.

The information in the hazard module is then fed to the vulnerability module to compute the damage to exposures given each event and the associated hazard intensity. The vulnerability module is usually consisted of multiple vulnerability curves, each of which corresponds to specific characteristics of the exposures and risk. One of the most common output of the vulnerability module is a damage ratio, that represents the ratio of the cost to repair an exposure to the cost of rebuilding it, at a given location under the associated risk conditions. It is possible that even identical buildings may experience different damages when being hit by the same hazard intensity due to the small differences and local building-specific effects. Therefore rather than a single point value, a distribution of possible damage ratios are usually provided for the vulnerability. Figure C.1 presents an example of vulnerability curve for hurricane with uncertain damage ratios given specific wind speed, along with the possible losses.

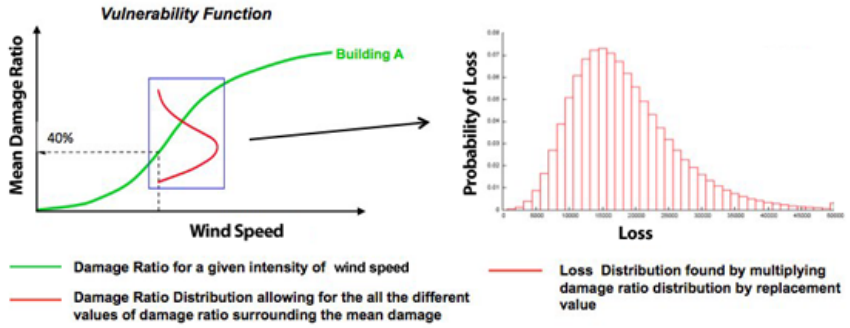


Figure C.1: Vulnerability function for a building depending on the wind speed (left) and the translation into loss distribution (right). Figure source: *Quantifying the Risk of Natural Catastrophes* by Shane Latchman at http://understandinguncertainty.org/node/622.

Financial module is responsible for the calculation of the loss distribution of an insured portfolio. It usually starts with a calculation of the ground up loss (GUL) to individual policyholder or exposure, and then applies the specific policy and programme level conditions such as limits, deductibles, and special conditions that

have been coded into the model to calculate the more complicated actual financial losses to the (re)insurers. The output is an event loss table (ELT) that provides the financial risk exposure to individual events. The table can be used to calculate the exceedance probability (EP) curves to conduct further assessment of the entire risk. An EP curve measures the probability of a given financial loss level being exceeded. Most commonly used EP curves include occurrence exceedance probability (OEP) and aggregate exceedance probability (AEP). The OEP tells the probability that there is any single event with a particular loss level or greater while AEP represents the probability of the total annual losses of all the possible events exceeding a particular loss level. Figure C.2 displays an example of the EP curve which could be AEP or OEP. The curve is plotted with loss against the so-called "return period" which is inversely related to the exceedance probability. Such representation is in fact an interpretation of the probability which is directly related to the simulated events and the associated simulated years. For example, the return period of 4,000 in Figure C.2 corresponds to the loss of around 20 million pounds. This is implying that the probability that a loss exceeds 20 million pounds in any given year is $1/4000 \approx 0.025\%$. We could also interpret such loss as a "1 in 4,000 years" event. Note that the "return period" representation and the simulated years, e.g. $1 \sim$ 10,000 years in Figure C.2, are just representation of the possible scenarios in a single year in the future and they are not what might happen in the next 10,000 years. When we consider the exceedance probability over a few years, the probability needs to be compounded. For example, if the annual exceedance probability of 1 million pounds in any given year is 0.1%, then the compounded exceedance probability of 5 years is roughly $1 - (1 - 0.1\%)^5 \approx 0.499\%$.

C.3 Oasis platform and ktools

Oasis is a not-for-profit platform to create and foster links throughout the wide community across business, academia and government. It provides an open marketplace for models and data leading to wider tools for catastrophe risk assessment. It was established in 2012 with support from the Insurance Industry, Innovate UK and

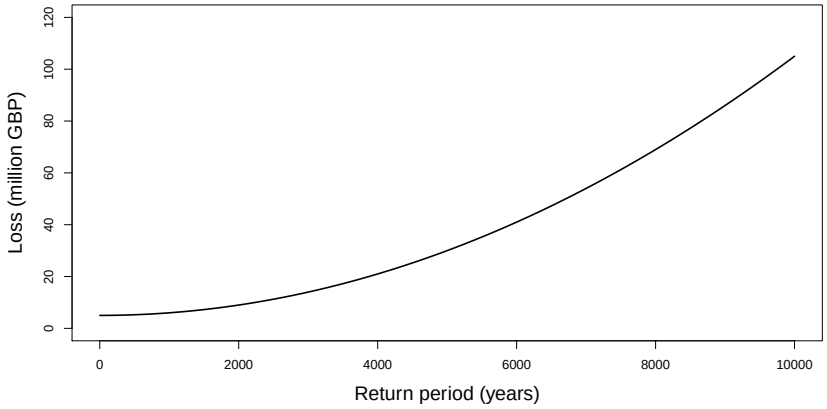


Figure C.2: An example of EP curve.

Climate KIC. It has been growing fast and produced an agnostic "plug and play" kernel that is very fast and flexible. Figure C.3 presents the diagram of the Oasis computational framework.

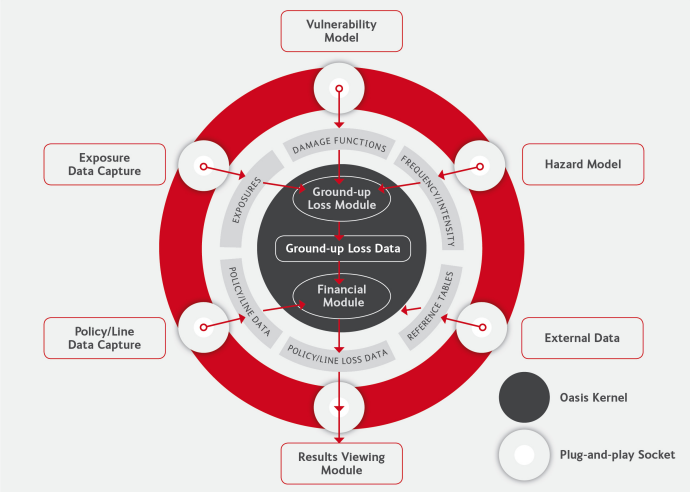


Figure C.3: Computational framework of OASIS platform.

In the core of the framework is the Oasis Kernel. The Kernel sits agnostically behind the "plug and play sockets" or connectors that relate the external actualised model and business data to the abstract Oasis computing structures such as Oasis Monte Carlo sampling of damage calculations. Oasis has developed the Kernel in

various environments and architectures. Here we employ the recently developed in-memory solution for the kernel which is called the kernel tools or ktools. It is written in C++ and C to provide streamed calculation at a high computational performance. The Kernel is provided as a toolkit of components which can be invoked at the user's convenience. Each component is a separately compiled executable with a binary data stream of inputs and outputs. The principle is to stream data through by event end-to-end, with multiple processes being used either sequentially or concurrently, at the control of the user. There is an implementation named "Reference Model" which can then be adapted for particular models or business needs. The workflow of the Reference Model is displayed in Figure C.4.

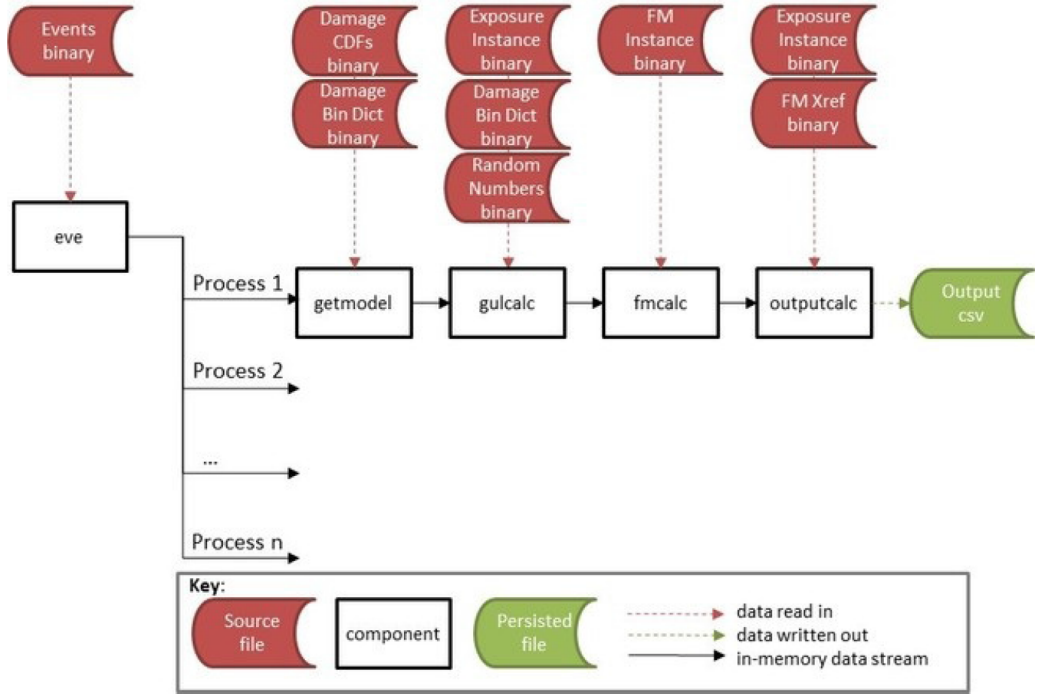


Figure C.4: Workflow and stream of ktools components.

We can see that there are several core components in the workflow. In particular, *eve* is the event distributing utility. It outputs subsets of the events as streams into the next component *getmodel*, based on the number of events in the input and the number of processes specified as a parameter. Then *getmodel* generates a stream of effective damageability cumulative distribution functions (CDFs) for each of the input streams of events. The output CDFs can be streamed into the next component

gulcalc or output to a file. The component *gulcalc* performs the ground up loss calculations using Monte Carlo sampling and numerical integration. The output is the Oasis kernel GUL sample table. This can be output to an external file or streamed into *fmcalc* or *outputcalc*. The *fmcalc* component calculates the insured loss based on the GUL samples and insured portfolio/programme descriptions. The output is the Oasis format loss sample table of the insured losses. The result can be streamed into *outputcalc* or output to a file. The component *outputcalc* is for output analysis on the GUL samples or insured loss samples. In the Reference Model, it is an ELT containing total insured value (TIV), sample mean and standard deviation of the losses for each event at the portfolio/programme summary level. The results are written directly into files as it is the end of the stream.

C.4 Synthetic studies with the UCL Cascadia tsunami hazard model

The UCL Cascadia tsunami hazard model is mainly developed by Serge Guillas, Simon Day and Andria Sarri. I was involved in the project at the initial stage where my duties are described previously in Appendix B. The UCL Cascadia tsunami hazard model incorporates novel coseismic characterisation of the potential tsunami sources with high-resolution tsunami simulations using VOLNA. There are 500 events across 43,826 area perils with non-zero inundation depths in at least one of the 500 events along the west coast of North America in the current version. More details about the hazard model can be found in Sarri (2015). Figure C.5 presents an example for the inundation over the whole study region produced by one of the 500 events in the UCL Cascadia tsunami hazard model.

We apply the Oasis ktools to the UCL Cascadia tsunami hazard model. The aims of this study are to highlight the precision, efficiency and the treatment of uncertainties of the UCL Cascadia tsunami hazard model, as well as to illustrate and examine the use of the Oasis platform with ktools. Because of the restricted access to the real insured portfolios including the exposures data, insurance policies and programmes, we only perform GUL analysis on synthetic exposures. The associ-

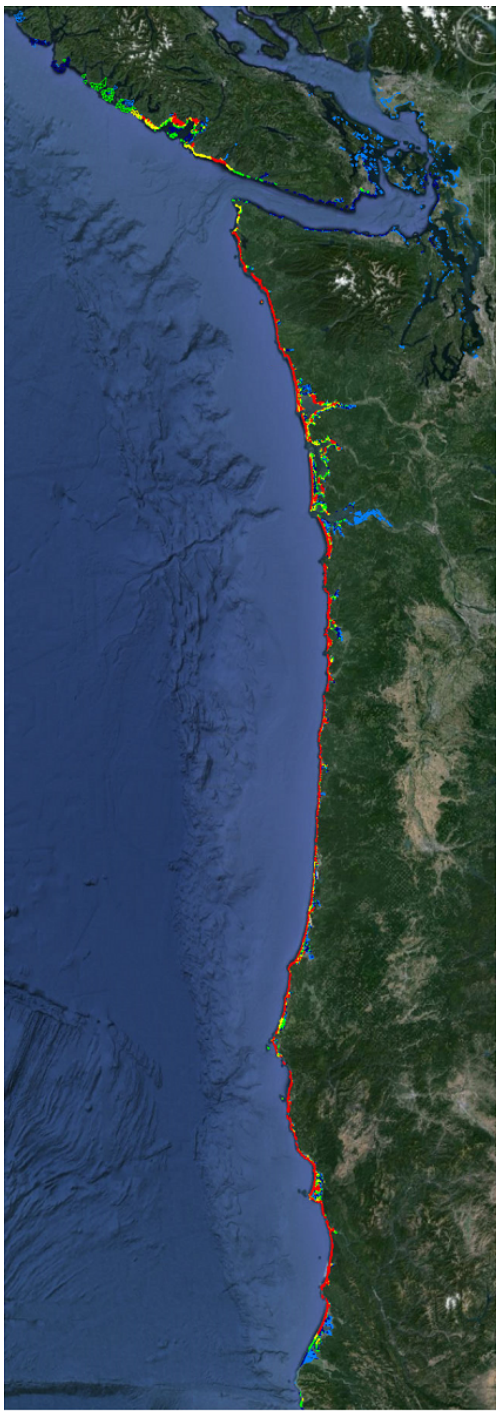


Figure C.5: Hazard intensity (inundation depth) produced by one of the 500 events in the UCL Cascadia tsunami hazard model over the whole region of west pacific coast of North America. The colours represent different levels of inundation depth with light blue for $0 \sim 2.5$ m, dark blue for $2.5 \sim 5$ m, green for $5 \sim 7.5$ m, yellow for $7.5 \sim 10$ m and red for 10+ m. Figure source: Sarri (2015).

ated components in the ktools are *eve*, *getmodel*, *gulcalc* and *outputcalc*. We are particularly interested in the performance of the core component *gulcalc* where the GUL samples are generated.

Before carrying out the studies, we need make some technical specifications. For the component *eve*, the event dictionary contains event ids from 1 to 500 for the 500 events. We also specify only one process in *eve*. To define the damage ratio, it is required to provide discretised damage bins rather than continuous curves because of the implementation of Oasis ktools. For example, the damage bins can be defined as $[0, 0]$, $[0, 0.1)$,..., $[0.9, 1)$, $[1, 1]$. The first and last bins allow the possibility that there is no or full damage in some cases. The damage CDFs contain the CDFs of each vulnerability curve, given the hazard intensity at each location for each event. These CDFs are also discretised according to the damage bins. The exposure data contains a list of exposures with their associated location id's corresponding to those in the hazard module, vulnerability id's representing the vulnerability curve that should be applied to the individual exposure, and total insured values.

C.4.1 Convergence and computing time of the GUL Monte Carlo sampling

In this study, we investigate the computing time and convergence of the GUL Monte Carlo sampling with respect to the discretisation of damage bins and vulnerability curves on a set of synthetic exposures. The simulations are run on the departmental server Speyburn at Statistical Science in UCL. We consider the following simulation specifications.

- **Exposure** The exposures are located at four towns: Victoria (BC, Canada), Aberdeen (Washington State, US), Long Beach (Washington State, US) and Pacific City (Oregon State, US). There are 550, 250, 100 and 100 insured buildings at the four regions respectively. The TIVs are sampled from Beta distribution between 0.1 $\sim$ 10 million US dollars. It is assumed that Victoria and Aberdeen have more expensive buildings than cheap buildings, while the other two regions have relatively less expensive buildings.

- **Damage bins** We consider four different discretisations for the damage bins: (1) $[0, 0]$, $[1, 1]$; (2) $[0, 0]$, $[0, 0.2)$, ..., $[0.8, 1)$, $[1, 1]$; (3) $[0, 0]$, $[0, 0.1)$, ..., $[0.9, 1)$, $[1, 1]$; (4) $[0, 0]$, $[0, 0.05)$, ..., $[0.95, 1)$, $[1, 1]$. Hence the number of bins are 2, 7, 12 and 22 respectively. Note that, the damage CDFs will also be discretised accordingly with each set of damage bins. For example, given the damage bins (3), the CDFs are defined at $0, 0.1, ..., 1$.

- **Vulnerability** Three vulnerability curves that are presented as the relationship between damage ratio (R) and hazard intensity are considered. In the current version of UCL Cascadia tsunami hazard model, the hazard intensity is represented with the inundation depth namely D.

  (1) Binary vulnerability function (Suppasri et al., 2011): $P(R = 1|D) = p(D)$ and $P(R = 0|D) = 1 - p(D)$. For $D \in (0, 40)$, $p(D) = \Phi(\frac{\log(D)-\mu}{\sigma})$, $\mu = 0.917$, $\sigma = 0.642$ where $\Phi(\cdot)$ is the CDF for standard Gaussian distribution, and $p(D = 0) = 0$, $p(D \geq 40) = 1$. Therefore, the building will be either completely undamaged or fully damaged with some probability depending on the hazard intensity.

  (2) Continuous vulnerability function 1: $P(R = 0|D = 0) = 1$, $P(R = 1|D \geq 40) = 1$, and for $D \in (0, 40)$, $P(R < r|HI) = \Phi(\frac{r-\mu'}{\sigma'})$, where $\mu' = \Phi(\frac{\log(D)-\mu}{\sigma})$, $\sigma' = 0.1$, $\mu = 0.917$ and $\sigma = 0.642$, $P(R = 0|D) = P(R < 0|D)$, $P(R = 1|D) = 1 - P(R < 1|D)$. That is to say the building will be completely undamaged when the inundation depth is zero or fully damaged when the inundation depth is equal to or greater than 40 m. Given a hazard inundation within $(0, 40)$, the damage ratio R can be any value within $[0, 1]$ following some zero-one-inflated truncated normal distribution: suppose there is a random variable $R' \sim N(\mu', \sigma')$, the damage ratio R is

$$R = \begin{cases} 0, & R' \leq 0 \\ R', & 0 < R' < 1 \\ 1, & R' \geq 1 \end{cases}.$$

(3) Continuous vulnerability function 2: This vulnerability function is the same as the vulnerability function (2) except $\sigma' = 0.2$. This implies that given a specific hazard intensity D, the variation of possible damage ratio is larger than that of the vulnerability function (2).

In the *gulcalc* process, Monte Carlo samples of potential damage are drawn according to the damage CDFs. The number of samples relates directly to the computational cost of this process as well as the convergence. We consider 35 increasing Monte Carlo sample size N from 10 to 5000 here: 10, 20, ..., 100, 200, 400, ..., 5000. Each set of N samples generates N possible loss values for each event, which is denoted by L_{ij} for event i and sample j. Then we can summarise those samples into the ELT of sample mean and standard deviation that are respectively $m_N(L_i) = \frac{1}{N}\sum_{j=1}^{N}(L_{ij})$ and $s_N(L_i) = [\frac{1}{N}\sum_{j=1}^{N}(L_{ij} - m(L_i))^2]^{1/2}$ for event $i = 1, .., 500$. For each event i, letting $m_{5000}(L_i)$ be the baseline, we can assess the convergence of the sample means in terms of relative errors $RE_N(L_i) = |L_N(L_i) - L_{5000}(L_i)|/|L_{5000}(L_i)|$ for $N < 5000$. Since there are 500 events each of which produces a relative error, we measure the overall convergence using the average relative errors $ARE = \frac{1}{N}\sum_{i=1}^{N} RE_N(L_i)$; see Figure C.6. In all cases, the relative errors decreases as the sample size increases. When there are 2 to 12 damage bins with the vulnerability function (1), more damage bins do not affect the description of a binary vulnerability and the lines coincide. In general, the errors for vulnerability function (2) are less than the other two due to its smaller variation in the possible damage ratios given a hazard intensity.

Note that the damage CDFs are discretised at damage bins. Figure C.6 only demonstrates the convergence of the Monte Carlo samples with specific damage bins and the associated discretised vulnerability curve. However, it doesn't necessarily indicate the convergence towards the true loss distribution of each vulnerability curve because of the approximation errors to the true vulnerability curve with finite discretisation. For the binary vulnerability (1), two damage bins are enough for such Bernoulli distribution. However, for the continuous vulnerability functions (2) and (3), a small number of damage bins may not be enough to describe the

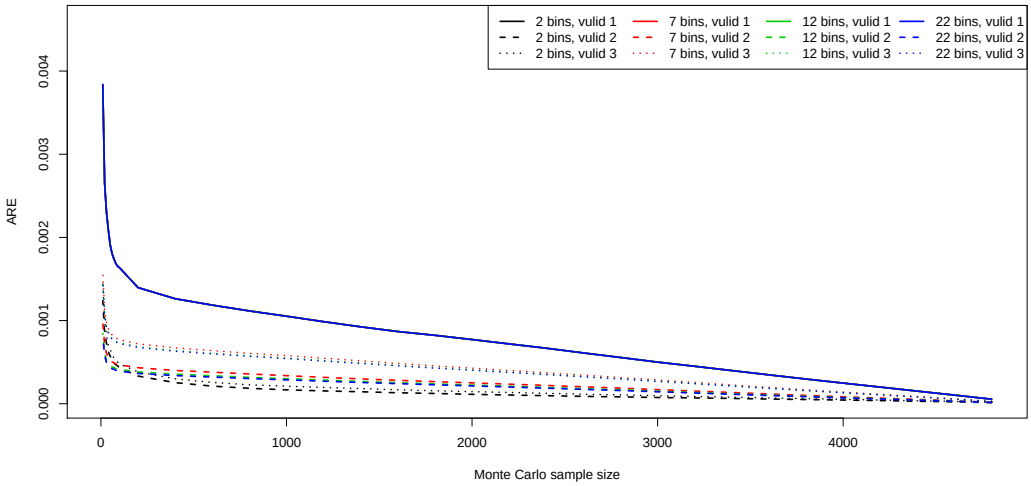


Figure C.6: AREs of sample mean losses using different Monte Carlo sample sizes against those using 5000 Monte Carlo samples. The legend "vulid" represents the type of vulnerability function, for example "vulid 1" corresponds to the vulnerability function (1).

damage distribution very well, thus result in misleading event losses. Therefore, the errors come from both Monte Carlo sampling and discretisation of the CDFs. We now assess the errors by looking at the convergence of the sample means to the respect analytic expectation of losses of the vulnerability functions. The results are displayed in Figure C.7. The large AREs for vulnerability functions (2) and (3) with 2 damage bins in the left panel suggest clearly that they never to converge to the respective expected losses. From the zoomed version for the other cases in the right panel, we can see that the convergence results are the same for the binary vulnerability functions (1) with 2 or more damage bins. But for the other two vulnerability functions, the errors decrease as there are more damage bins to describe the CDFs.

Figure C.8 displays the run time elapsed for calling the two components *gulcalc* and *outputcalc* to do the Monte Carlo sampling of possible damages and to summarise the samples into event loss tables. In each case, the computing time increases linearly with the Monte Carlo sample size in general. For fixed number of Monte Carlo samples, the computing time takes longer when there are more damage bins. The sampling procedure seems to take more time for vulnerability function (3) than the other two and it also takes longer for vulnerability function (2)

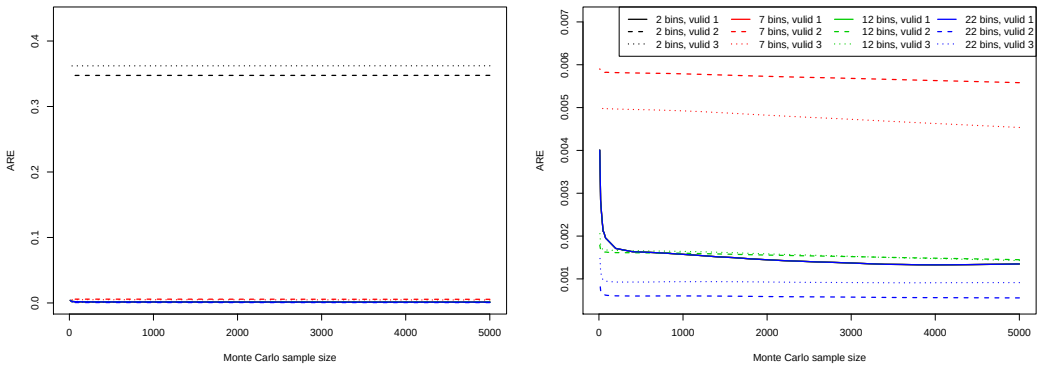


Figure C.7: AREs of sample mean losses using different Monte Carlo sample sizes against the corresponding analytic expectations for three vulnerability functions (vulid 1, 2 and 3) and different discretisation of damage bins. Right panel zooms the lower part of left panel.

than (1) in most cases, that is consistent with the increasing complexity of the three vulnerability functions.

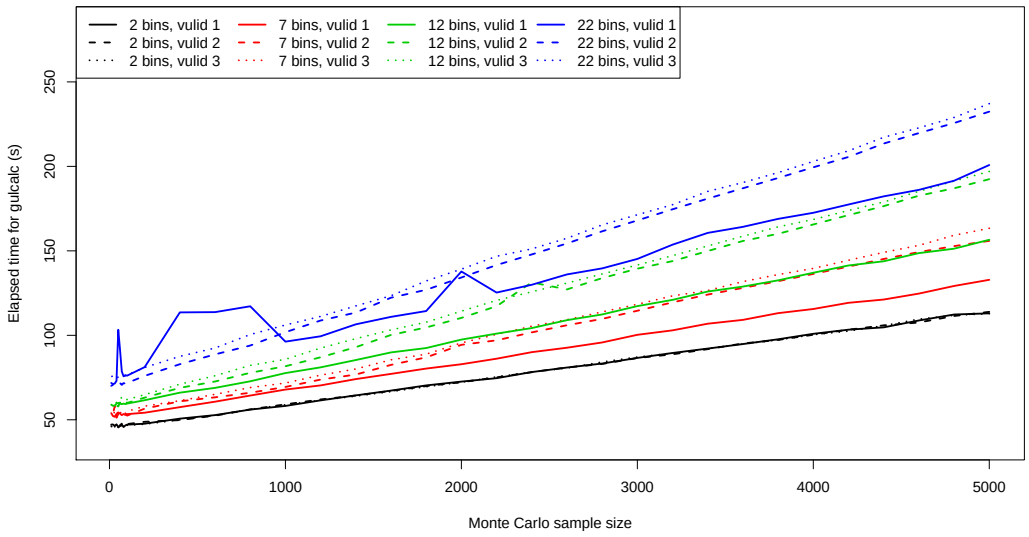


Figure C.8: Elapsed time for running *gulcalc* and *outputcalc* with increasing Monte Carlo sample size for three vulnerability functions (vulid 1, 2 and 3) and different discretisation of damage bins.

C.4.2 Sensitivity of computing time to the number of exposures

The number of exposures might also relate to the computing time of GUL sampling. We randomly generate n buildings with random exclusive locations and investigate

the associated computing time of *gulcalc* and *outputcalc*. For illustration, we consider $n = 1000, 2000, ..., 20000$ and the binary vulnerability function (1) with two damage bins. Figure C.9 presents the computing time, with three different Monte Carlo sample sizes of 100, 1000 and 5000, against the increasing number of exposure buildings. It is clear that the computing time grows linearly as there are more exposures to be considered. The slope of such growth in the computing time also increases with the number of Monte Carlo samples.

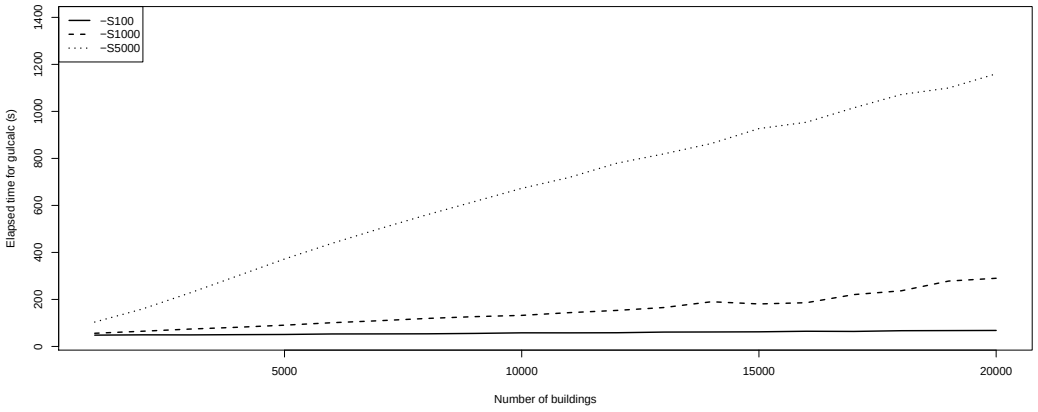


Figure C.9: Elapsed time for running *gulcalc* and *outputcalc* with increasing number of exposures using Monte Carlo sample size of 100, 1000 and 5000.

C.4.3 Computing time comparison between R and ktools

In this section, we compare the computing time used to generate GUL samples and summarise into ELTs using ktools (*gulcalc* and *outputcalc*) and using the equivalents in R. The binary vulnerability (1) with various number of Monte Carlo sampling sizes and the exposure set as described in Section C.4.1 are applied. In R, we mimic the same procedure by looping over each exposure and event and generating the GUL samples. Then the samples are summarised to sample mean and standard deviation as in the ktools component *outputcalc* and output to a file of the same format as in ktools. Figure C.10 compares the computing time for the procedure using ktools and R respectively. In general, the computing time of ktools starts with higher values but increases much slower than R. The higher starting values of ktools is due to the initial process of the component *gulcalc* such as reading in CDFs from

binary file and matching the exposures with their corresponding CDFs, that are not included in the R equivalent. The slower increasing slope using ktools indicates clearly that ktools should be more efficient than the R equivalent.

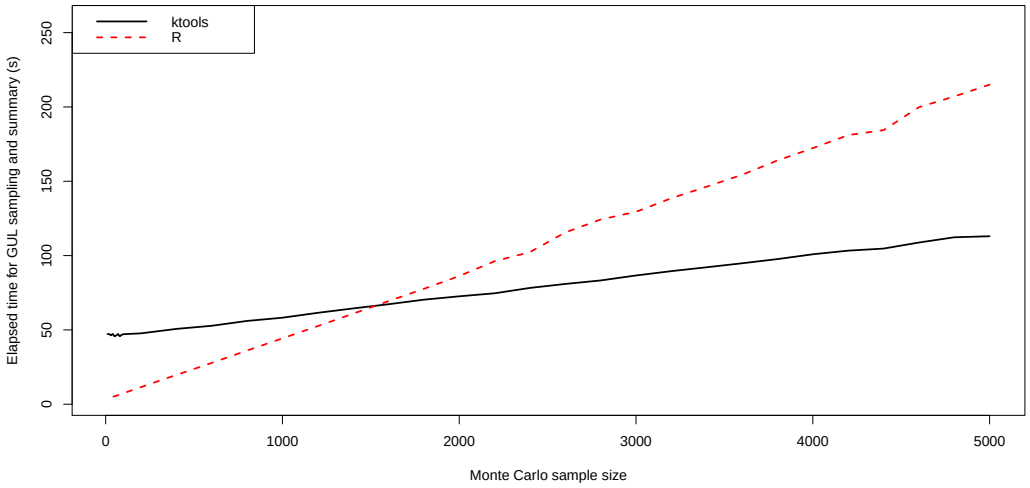


Figure C.10: Elapsed time for GUL sampling and summary with various Monte Carlo sample sizes using ktools and R.

C.4.4 Realistic portfolio illustration

In this section, we illustrate the UCL Cascadia tsunami hazard with a realistic worldwide portfolio that is shared by Oasis. Though realistic, only eight exposures in this portfolio are covered by the UCL Cascadia tsunami hazard model. The locations of these eight exposures and their respective index (areaperil_id) in the UCL Cascadia tsunami hazard model are displayed in Table C.1.

In the portfolio, the same insured value is assigned to all the exposures. Hence, we set TIV to 1 unit without loss in generality. Firstly, we investigate the convergence of the sample mean and standard deviation using Monte Carlo sampling to the analytic mean and standard deviation of the GULs for the 500 events. The binary vulnerability function (1) is employed. The average relative errors are presented in Figure C.11. It is clear that both sample means and standard deviations converge to the analytic values as the Monte Carlo sample size increases.

The EP curve for the GULs of these 500 events is shown in Figure C.12. Note

Table C.1: Locations of the exposures and the associated areaperil index in the UCL Cascadia tsunami hazard model.

item_id	areaperil_id	latitude	longitude
1	20200	46.1907	−123.968
2	18025	46.1326	−123.922
3	8858	47.0865	−124.113
4	22832	46.9859	−124.156
5	13442	46.8833	−124.110
6	10308	46.3305	−124.010
7	14633	46.3872	−124.040
8	21805	46.5424	−124.046

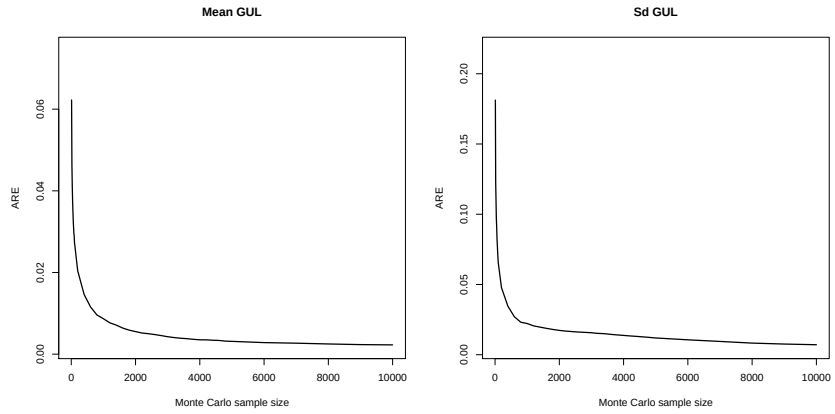


Figure C.11: AREs of GUL sample means and standard deviations against the analytic values for the ISCM portfolio.

that we assume these 500 possible events happen in 500 pseudo-years period with one single event occurs per year. This corresponds to the frequency rate of one event per year which is obviously unrealistic while the real tsunami events are actually very rare. However, we just make this assumption to use the EP curve as a graphical tool to present losses of the 500 events. The curve could be easily scaled to accommodate more realistic frequency assumption using more elegant statistical and probabilistic models such as Poisson processes. We can see that the 500 events in the UCL Cascadia tsunami hazard model could result in losses from 2 to 6.5 units roughly and the standard deviations suggest significant uncertainty in the losses due to the random nature of the vulnerability function.

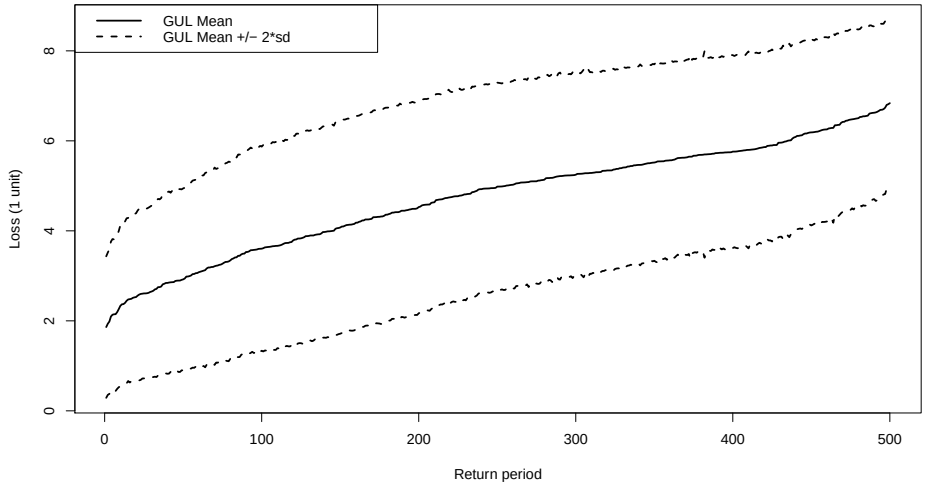


Figure C.12: EP curves for the ISCM portfolio.

C.5 Impact of the uncertainties in the bathymetry on GULs

In this section, we highlight the financial impact of the uncertainties in the bathymetry on synthetic insurance portfolios through GUL analysis. We have demonstrated that tsunami wave heights could be affected by the uncertainties in the bathymetry as shown in Figure 3.9. These variations can be propagated to the damage to coastal structures and hence the financial losses which are of great interest to the insurance industry.

We employ the synthetic tsunami case study in Section 3.5.2. Here we define the tsunami events with a single parameter h_{max} and consider the bathymetry as the primary uncertainties in the events. To highlight the impact of the uncertainties in the bathymetry, we compare two scenarios: (1) bathymetry are assumed to be well-known and fixed at the posterior means; (2) bathymetry are assumed to be uncertain. In the first scenario, for each event that is defined with a specific h_{max}, the hazard intensity (the inundation depth in this context) is certain which could lead to some damage to properties according to the assumed vulnerability function. In the second scenario, the hazard intensity for each event is random instead of a fixed value, which introduces a layer of uncertainties into the damage. We apply the

binary vulnerability function as described in Appendix C.4.1. This binary vulnerability function implies that given a specific hazard intensity, the damage ratio to the property is either 1 or 0 with some probability depending on the hazard intensity. This brings another layer of uncertainties into the damages which is considered to be the secondary uncertainty. Therefore, in the first scenario, the secondary uncertainty associated with the vulnerability function is the only source of uncertainty in the resulting losses, while both primary uncertainty from the bathymetry and the secondary uncertainty from the vulnerability function contribute to the overall uncertainty in the final losses in the second scenario. In this study, we do not present the secondary uncertainty to highlight the impact of the primary uncertainty associated with the bathymetry. We simulate 300 tsunami events with 300 random samples for h_{max} drawn from a Normal distribution $N(3, 1)$ truncated at 0 and 5; see in Figure C.13 a histogram for these samples. These h_{max}'s represent various intensities of possible seabed displacement that relate to the general scales of the tsunami waves.

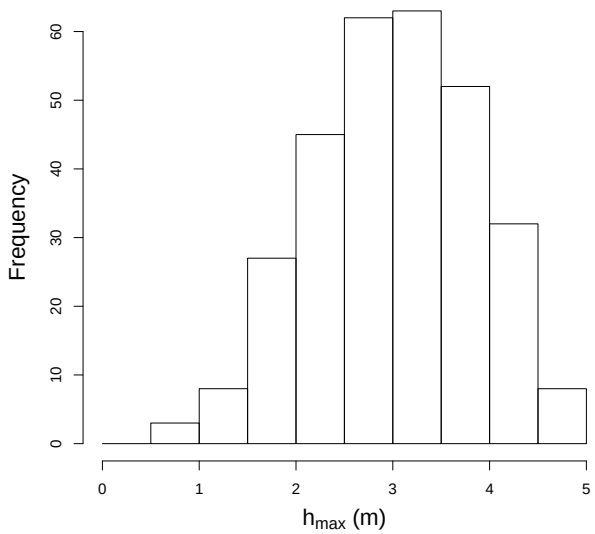


Figure C.13: Histogram of 300 random samples of h_{max} that generate 300 tsunami events.

The peril area $(s_1, s_2) \in [100, 9900] \times [-3000, 3000]$ are considered. The whole area is divided into 49 cells according to the value of the first coordinate s_1,

i.e. $[100, 300)$, $[300, 500)$,..., $[9700, 9900]$. The hazard intensity is assumed to be the same over each cell regardless of different locations. This discretisation does not depend on the s_2 coordinate based on the assumption that the true bathymetry is only dependent on s_1. Note that the random samples and posterior means of bathymetry deviate from the true bathymetry and are likely to be varying along s_2 even for the same s_1, hence the resulting inundation depths are likely to be different at different s_2 for the same s_1. For illustration purpose and simplicity, we do not consider the variations along the s_2 coordinate here. To construct a synthetic insured portfolio, we draw 1000 locations over the peril area uniformly to represent the exposures. The associated total insured values (TIVs) are sampled uniformly between 0.5 and 10 million US dollars. Figure C.14 presents the locations of these simulated exposures with the associated TIVs.

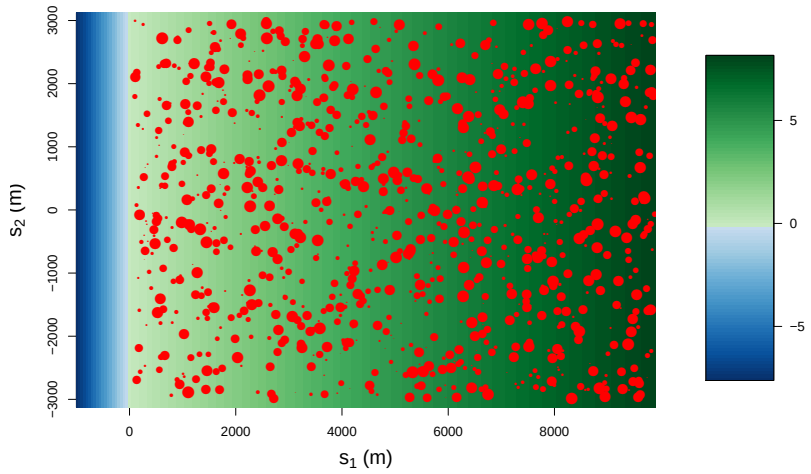


Figure C.14: Locations of exposures in the synthetic portfolio; size proportional to the associated TIVs ranging from 0.5 to 10 million USD.

We are able to produce the hazard map of each tsunami event by running the tsunami code VOLNA with a few well-designed simulations and applying the statistical emulation to predict more scenarios. Figure C.15 presents the hazard maps that are obtained using the posterior mean bathymetry at two survey levels for the tsunami event when $h_{max} = 3.50$ m. The hazard maps are certain since there are no uncertainties in the inundation. For comparison, we also produce another sets

of hazard maps for the same tsunami event by taking into account the uncertainties in the bathymetry. In particular, for each tsunami event, we predict the inundation depths over the peril area with 1000 random samples from the posterior of bathymetry surface using the emulation technique in Chapter 3. Therefore, the hazard intensity at any location for a specific tsunami event is uncertain. Figure C.16 shows the mean and standard deviation of the inundation depths over the peril area for the same tsunami event as in Figure C.15 when $h_{max} = 3.50$ m. It is clear that the mean hazard map are similar to those with fixed bathymetry. However, the significant variations in the hazard intensity, especially at the near-shore area, due to the uncertainties in the bathymetry are highlighted with the standard deviation map.

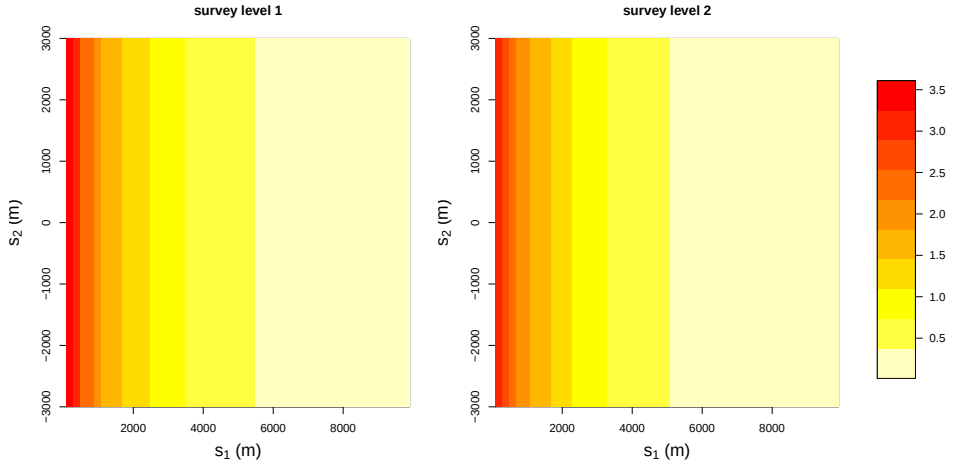


Figure C.15: Hazard intensity (inundation depth in metres) over the peril using the fixed mean bathymetry at two survey levels for tsunami event when $h_{max} = 3.50$ m.

The binary vulnerability function is employed such that given a fixed inundation depth D, the property is either damaged completely or undamaged. The damage ratio R is defined through a Bernoulli distribution that $P(R = 1|D) = p(D)$ and $P(R = 0|D) = 1 - p(D)$ with

$$p(D) = \begin{cases} 0, & D = 0 \\ \Phi(\frac{log(D)-\mu}{\sigma}), & 0 < D < 40 \\ 1, & D \geq 40 \end{cases},$$

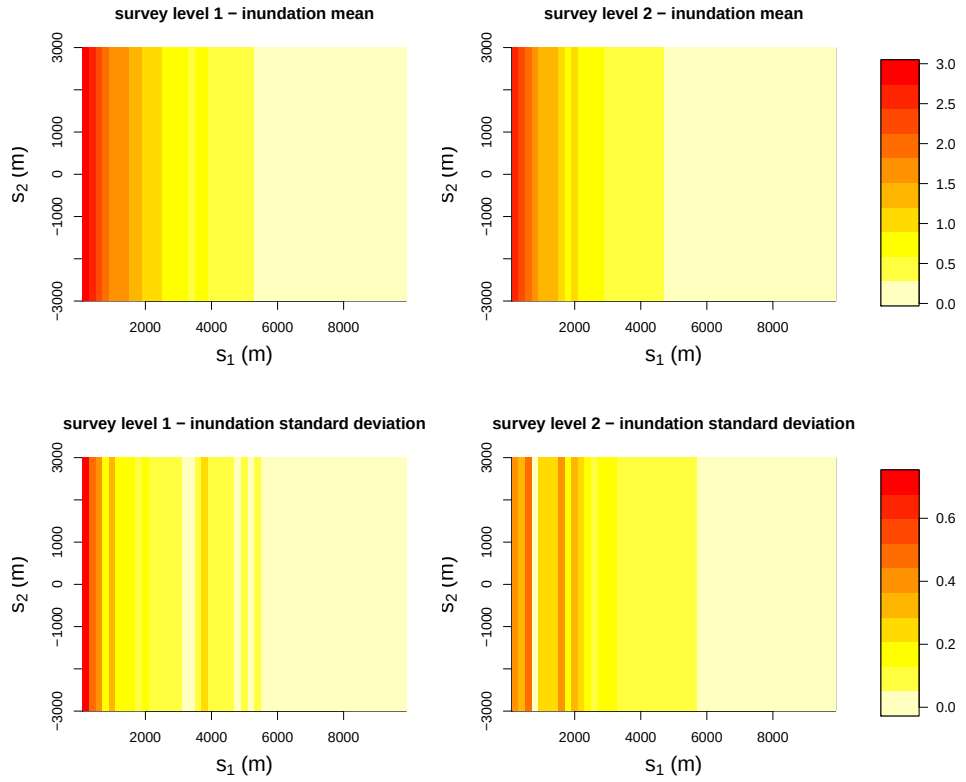


Figure C.16: Mean and standard deviation of the hazard intensity (inundation depth in metres) over the peril using the uncertain bathymetry samples at two survey levels for tsunami event when $h_{max} = 3.50$ m.

where $\mu = 0.917$, $\sigma = 0.642$ and $\Phi(\cdot)$ is the CDF of standard Gaussian distribution. Then, we are able to calculate the expected total ground up losses for each event through the Oasis platform. The EP curves of these 300 possible events are presented in Figure C.17. The curves here only serve as an overview of the ground up losses generated by the possible 300 events, hence are plotted simply against a return period of $1 \sim 300$ years. The variations in the EP curves are clearly significant when taking into account the uncertainties in the bathymetry. The difference between the potential losses using fixed bathymetry and those using uncertain bathymetry could range around -700 $\sim$ 100 million dollars. These variations provide additional insights to insurance companies to make decisions such as pricing their policies, purchasing re-insurance or allocating their capital.

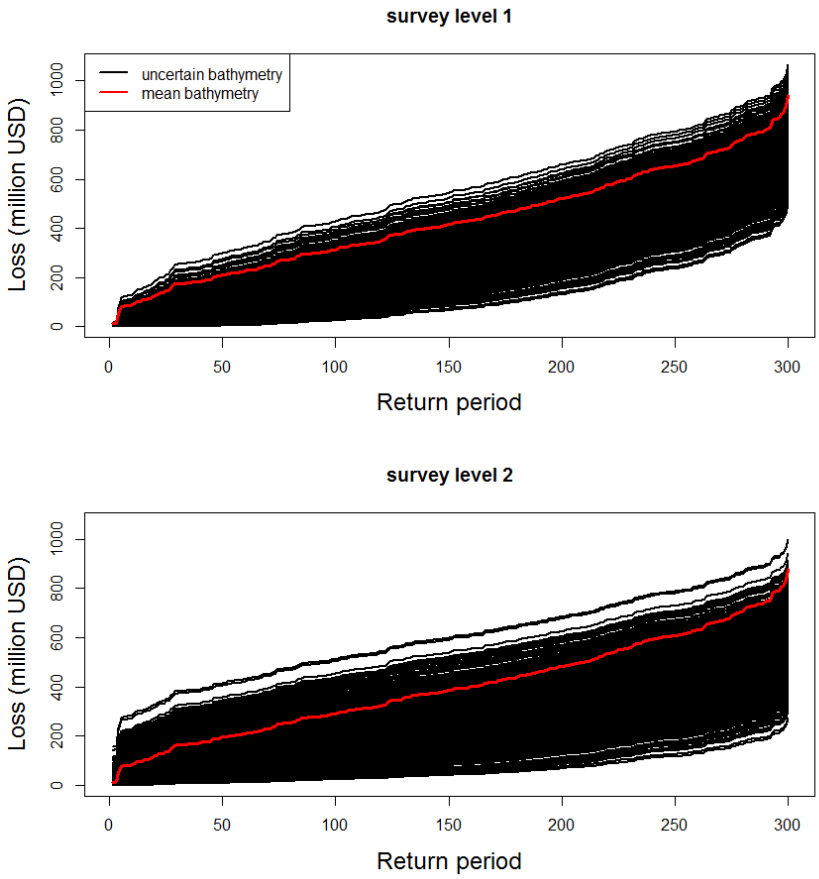


Figure C.17: EP curves of the 300 tsunami events with fixed mean bathymetry or uncertain bathymetry.

Bibliography

Adragni, K. P. and Raim, A. (2014). ldr: an R software package for likelihood-based sufficient dimension reduction. *Journal of Statistical Software*, 61(3):1–21.

Amante, C. and Eakins, B. W. (2009). *ETOPO1 1 arc-minute global relief model: procedures, data sources and analysis*. US Department of Commerce, National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service, National Geophysical Data Center, Marine Geology and Geophysics Division.

Amini, A. A. and Wainwright, M. J. (2012). Sampled forms of functional PCA in reproducing kernel Hilbert spaces. *The Annals of Statistics*, 40(5):2483–2510.

Andrianakis, I. and Challenor, P. G. (2012). The effect of the nugget on Gaussian process emulators of computer models. *Computational Statistics & Data Analysis*, 56(12):4215–4228.

Awanou, G., Lai, M.-J., and Wenston, P. (2006). The multivariate spline method for scattered data fitting and numerical solutions of partial differential equations. *Wavelets and splines: Athens*, pages 24–74.

Babuska, I., Szabo, B. A., and Katz, I. N. (1981). The p-version of the finite element method. *SIAM Journal on Numerical Analysis*, 18(3):515–545.

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004). *Hierarchical modeling and analysis for spatial data*. CRC Press.

Beck, J. and Guillas, S. (2016). Sequential design with mutual information for computer experiments (MICE): emulation of a tsunami model. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):739–766.

Becker, J., Sandwell, D., Smith, W., Braud, J., Binder, B., Depner, J., Fabre, D., Factor, J., Ingalls, S., Kim, S., et al. (2009). Global bathymetry and elevation data at 30 arc seconds resolution: SRTM30_PLUS. *Marine Geodesy*, 32(4):355–371.

Bolin, D. and Lindgren, F. (2013). A comparison between Markov approximations and other methods for large spatial data sets. *Computational Statistics & Data Analysis*, 61:7–21.

Brenner, S. C. and Scott, L. R. (2008). *The mathematical theory of finite element methods*, volume 15. Springer.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, 97(2):109–131.

Chen, C. M. and Thomée, V. (1985). The lumped mass finite element method for a parabolic problem. *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, 26(03):329–354.

Constantine, P. and Gleich, D. (2015). Computing active subspaces with Monte Carlo. *arXiv preprint arXiv:1408.0545*.

Constantine, P. G., Dow, E., and Wang, Q. (2014). Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425):177–189.

Cook, R. D. (2007). Fisher lecture: dimension reduction in regression. *Statistical Science*, pages 1–26.

Cook, R. D. (2009). *Regression graphics: ideas for studying regressions through graphics*, volume 482. John Wiley & Sons.

Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4):485–501.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction" by Li, K.C. *Journal of the American Statistical Association*, 86(414):328–332.

Cressie, N. (1993). *Statistics for spatial data*, volume 900. Wiley New York.

Diggle, P. and Ribeiro, P. J. (2007). *Model-based geostatistics*. Springer.

Diggle, P. J., Menezes, R., and Su, T.-L. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232.

Dutykh, D., Poncet, R., and Dias, F. (2011). The VOLNA code for the numerical modeling of tsunami waves: generation, propagation and inundation. *European Journal of Mechanics-B/Fluids*, 30(6):598–615.

Eakins, B. W. and Taylor, L. A. (2010). *Seamlessly integrating bathymetric and topographic data to support tsunami modeling and forecasting efforts*, chapter 2. ESRI Press, Relands.

Elliott, M. R. and Davis, W. W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):595–609.

Ettinger, B., Guillas, S., and Lai, M.-J. (2012). Bivariate splines for ozone concentration forecasting. *Environmetrics*, 23(4):317–328.

Fasshauer, G. E. (2011). Positive definite kernels: past, present and future. *Dolomite Research Notes on Approximation*, 4:21–63.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *The Journal of Machine Learning Research*, 5:73–99.

Fukumizu, K. and Leng, C. (2014). Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, 109(505):359–370.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Giles, M. B. (2008). Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617.

Giorgi, E., Sesay, S. S., Terlouw, D. J., and Diggle, P. J. (2015). Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(2):445–464.

Goldstein, M. and Wooff, D. (2007). *Bayes linear statistics: theory and methods*, volume 716. John Wiley & Sons.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

González, F. I., Geist, E. L., Jaffe, B., Kânoğlu, U., Mofjeld, H., Synolakis, C. E., Titov, V. V., Arcas, D., Bellomo, D., Carlton, D., et al. (2009). Probabilistic tsunami hazard assessment at Seaside, Oregon, for near-and far-field seismic sources. *Journal of Geophysical Research: Oceans (1978–2012)*, 114(C11).

Goto, C., Ogawa, Y., Shuto, N., and Imamura, F. (1997). *IUGG/IOC Time Project: Numerical method of tsunami simulation with the leap-frog scheme*. Technical report, UNESCO.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Guillas, S. and Lai, M.-J. (2010). Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics*, 22(4):477–497.

Hall, J. K. (2006). GEBCO centennial special issue – charting the secret world of the ocean floor: the GEBCO project 1903–2003. *Marine Geophysical Researches*, 27(1):1–5.

Hell, B. and Jakobsson, M. (2011). Gridding heterogeneous bathymetric data sets with stacked continuous curvature splines in tension. *Marine Geophysical Research*, 32(4):493–501.

Henderson, D. A., Boys, R. J., Krishnan, K. J., Lawless, C., and Wilkinson, D. J. (2009). Bayesian emulation and calibration of a stochastic computer model of mitochondrial DNA deletions in substantia nigra neurons. *Journal of the American Statistical Association*, 104(485):76–87.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Hu, X.-L., Han, D.-F., and Lai, M.-J. (2007). Bivariate splines of various degrees for numerical solution of partial differential equations. *SIAM Journal on Scientific Computing*, 29(3):1338–1354.

Iglesias, O., Lastras, G., Souto, C., Costa, S., and Canals, M. (2014). Effects of coastal submarine canyons on tsunami propagation and impact. *Marine Geology*, 350:39–51.

Imamura, F. (1996). *Long-wave runup models*, chapter Simulation of wave-packet propagation along sloping beach by TUNAMI-code, pages 231–241. World Scientific.

Iooss, B. and Ribatet, M. (2009). Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering & System Safety*, 94(7):1194–1204.

Kim, D. H. (2011). *Partial sufficient dimension reduction in regression*. PhD thesis, University of Minnesota.

Lai, M.-J. and Schumaker, L. L. (2007). *Spline functions on triangulations*, volume 110. Cambridge University Press.

Lai, M.-J., Shum, C. K., Baramidze, V., and Wenston, P. (2009). Triangulated spherical splines for geopotential reconstruction. *Journal of Geodesy*, 83(8):695–708.

Li, B., Wen, S., and Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association*, 103(483):1177–1186.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86(414):316–342.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Liu, P. L.-F., Woo, S.-B., and Cho, Y.-K. (1998). *Computer programs for tsunami propagation and inundation*. Technical report, Cornell University.

Liu, X. and Guillas, S. (2016). Dimension reduction for emulation: application to the influence of bathymetry on tsunami heights. *arXiv preprint arXiv:1603.07888*.

Liu, X., Guillas, S., and Lai, M.-J. (2015). Efficient spatial modelling using the SPDE approach with bivariate splines. *Journal of Computational and Graphical Statistics*, accepted.

Lohr, S. and Rao, J. K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101(475):1019–1030.

Manzi, G., Spiegelhalter, D. J., Turner, R. M., Flowers, J., and Thompson, S. G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(1):31–50.

Marrel, A., Iooss, B., Da Veiga, S., and Ribatet, M. (2012). Global sensitivity analysis of stochastic computer models with joint metamodels. *Statistics and Computing*, 22(3):833–847.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, 67:68–83.

Moriarity, C. and Scheuren, F. (2001). Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17(3):407.

NOAA NCEI (2016). *U.S. Coastal Relief Model*. NOAA National Centers for Environmental Information, http://www.ngdc.noaa.gov/mgg/coastal/crm.html.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599.

O'Hagan, A. (1994). *Kendall's advanced theory of statistics, Volume 2B: Bayesian inferenc*. Edward Arnold, London.

Paciorek, C. J. and Schervish, M. J. (2004). Nonstationary covariance functions for Gaussian process regression. In *Advances in Neural Information Processing Systems*, pages 273–280.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press.

Rue, H. and Held, L. (2004). *Gaussian Markov random fields: theory and applications*. CRC Press.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science*, pages 409–423.

Sangalli, L. M., Ramsay, J. O., and Ramsay, T. O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):681–703.

Sarri, A. (2015). *Emulators in the investigation of sensitivities and uncertainties in tsunami models*. PhD thesis, University College London.

Sarri, A., Guillas, S., and Dias, F. (2012). Statistical emulation of a tsunami model for sensitivity analysis and uncertainty quantification. *Natural Hazards and Earth System Science*, 12(6):2003–2018.

Simpson, D., Illian, J. B., Lindgren, F., Srbye, S. H., and Rue, H. (2016). Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1):49–70.

Simpson, D., Lindgren, F., and Rue, H. (2012). Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics*, 1:16–29.

Smith, W. H. F. and Wessel, P. (1990). Gridding with continuous curvature splines in tension. *Geophysics*, 55(3):293–305.

Snelson, E. and Ghahramani, Z. (2005). Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, pages 1257–1264.

Sraj, I., Mandli, K. T., Knio, O. M., Dawson, C. N., and Hoteit, I. (2014). Uncertainty quantification and inference of Manning's friction coefficients using dart buoy data during the Tōhoku tsunami. *Ocean Modelling*, 83:82–97.

Stefanescu, E. R., Bursik, M., Cordoba, G., Dalbey, K., Jones, M. D., Patra, A. K., Pieri, D. C., Pitman, E. B., and Sheridan, M. F. (2012a). Digital elevation model uncertainty and hazard analysis using a geophysical flow model. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 468(2142):1543–1563.

Stefanescu, E. R., Bursik, M., and Patra, A. K. (2012b). Effect of digital elevation model on Mohr-Coulomb geophysical flow model output. *Natural hazards*, 62(2):635–656.

Suppasri, A., Koshimura, S., and Imamura, F. (2011). Developing tsunami fragility curves based on the satellite remote sensing and the numerical modeling of the 2004 Indian Ocean tsunami in Thailand. *Natural Hazards and Earth System Sciences*, 11(1):173–189.

Titov, V. V. and Gonzalez, F. I. (1997). *Implementation and testing of the Method of Splitting Tsunami (MOST) model*. Technical Report ERL PMEL-112, US Department of Commerce, National Oceanic and Atmospheric Administration, Environmental Research Laboratories, Pacific Marine Environmental Laboratory.

Turner, R. M., Spiegelhalter, D. J., Smith, G., and Thompson, S. G. (2009). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):21–47.

Walsh, J. B. (1986). *An introduction to stochastic partial differential equations*. Springer.

Wessel, P. and Bercovici, D. (1998). Interpolation with splines in tension: a Green's function approach. *Mathematical Geology*, 30(1):77–93.

Wessel, P. and Smith, W. H. (1998). New, improved version of generic mapping tools released. *Eos, Transactions American Geophysical Union*, 79(47):579–579.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, pages 434–449.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2):975–994.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410.

Yin, X. and Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):879–892.